PB82-114364

Data Validation Conference Proceedings

(U.S.) Environmental Monitoring and Support Lab.
Research Triangle Park, NC

Sep 79

U.S. Department of Commerce
National Technical Information Service
NTIS

United States
Environmental Protection
Agency

Environmental Monitoring
Systems Laboratory
Research Triangle Park NC 27711

EPA-600/9-79-042
September 1979

PB82-114364

Research and Development

EPA

Data Validation Conference Proceedings

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the MISCELLANEOUS REPORTS series. This series is reserved for reports whose content does not fit into one of the other specific series. Conference proceedings, annual reports, and bibliographies are examples of miscellaneous reports.

EPA REVIEW NOTICE

This report has been reviewed by the U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policy of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/9-79-042	2.	3. RECIPIENT'S ACCESSION NO. PB82 114364
4. TITLE AND SUBTITLE DATA VALIDATION CONFERENCE, Proceedings		5. REPORT DATE September 1979 6. PERFORMING ORGANIZATION CODE PF U282024
7. AUTHOR(S) Raymond C. Rhodes and Seymour Hocheiser, Editors		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Office of Research and Development Environmental Monitoring and Support Laboratory Research Triangle Park, N. C. 27711		10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA 600/08
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

The proceedings document technical presentations made at a 1-day conference on Data Validation for environmental data. The conference was hosted and sponsored by the U.S. Environmental Protection Agency, Research Triangle Park Interlaboratory Quality Assurance Coordinating Committee on November 4, 1977, at the Research Triangle Park. Various approaches and techniques used for data validation are presented.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Data Validation Data Screening Data Editing Quality Assurance Outliers Statistics Environmental Data	Environmental monitoring Data management	43F 68A

18. DISTRIBUTION STATEMENT Release to public	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 315
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (9-73)

DATA VALIDATION CONFERENCE

Proceedings

Hosted and Sponsored by

The U.S. Environmental Protection Agency

RTP Interlaboratory Quality Assurance Coordinating Committee

November 4, 1977

Edited by

Raymond C. Rhodes

and

Seymour Hochheiser

Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

DISCLAIMER

This report is a compilation of the papers presented at the Data Validation Conference. Each individual paper may not have received peer technical review. Technical review and clearance of these proceedings was based primarily on the review of the executive summary and on the general merits of the proceedings as a total entity.

The content of these proceedings do not necessarily reflect the views and policies of the U.S. Environmental Protection Agency nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

FOREWORD

Measurement and monitoring research efforts are designed to anticipate potential environmental problems, to support regulatory actions by developing an in-depth understanding of the nature and processes that impact health and the ecology, to provide innovative means of monitoring compliance with regulations and to evaluate the effectiveness of health and environmental protection efforts through the monitoring of long-term trends. The Environmental Monitoring Systems Laboratory, Research Triangle Park, North Carolina, has the responsibility for: assessment of environmental monitoring technology and systems; implementation of agency-wide quality assurance programs for air pollution measurement systems; and supplying technical support to other groups in the Agency including the Office of Air, Noise and Radiation, the Office of Toxic Substances and the Office of Enforcement.

Data validation, an element of quality assurance is necessary to provide accurate and reliable environmental data. Data of known and acceptable quality are needed for measuring compliance with regulations, assessing health effects, and developing optimum strategies to cope with environmental pollution situations. A unified treatment of validation of particular types of data bases is needed to support broad-scale uses of these data. Current in-use data validation procedures were presented at the conference to promote a better understanding of available techniques. Hopefully, the conference and these proceedings will provide an impetus toward the development of more unified and systematic approaches to data validation.

Thomsas R. Hauser, Ph. D.
Director
Environmental Monitoring Systems Laboratory
Research Triangle Park, North Carolina

ABSTRACT

These proceedings are a record for future reference of the technical presentations made at a conference on Data Validation for environmental data. The conference was hosted and sponsored by the U. S. Environmental Protection Agency, Research Triangle Park Interlaboratory Quality Assurance Coordinating Committee on November 4, 1977, at the Research Triangle Park. Various data validation approaches and techniques were presented and are documented in this publication.

CONTENTS

FOREWORD iii

ACKNOWLEDGEMENTS vii

INTRODUCTION 1

EXECUTIVE SUMMARY AND RECOMMENDATIONS 3
Seymour Hochheiser
Raymond C. Rhodes

WHAT IS DATA VALIDATION? 7
Raymond C. Rhodes

THE SHEWHART CONTROL CHART TEST FOR SCREENING 24-HOUR AIR POLLUTION MEASUREMENTS 17
William F. Hunt

DISTRIBUTION GAP TEST FOR HOURLY AIR POLLUTION DATA 25
Thomas C. Curran

USE OF STATISTICAL SAMPLING IN VALIDATING HEALTH EFFECTS DATA 31
Carolyn P. Chamblee

USE OF SUCCESSIVE TIME DIFFERENCES AND DIXON RATIO TEST FOR DATA VALIDATION 39
Tyler Hartwell

CLUSTER ANALYSIS AS A DATA VALIDATION TECHNIQUE 71
Harold L. Crutcher

ENGINEERING COMPUTATIONS AND DATA COLLECTION FORMATS USEFUL IN DATA VALIDATION 81
A. Carl Nelson, Jr.

VALIDATION PROCEDURES APPLIED TO IN-USE MOTOR VEHICLE EMISSION DATA 99
Marcia E. Williams

DATA VALIDATION TECHNIQUES USED IN MOBILE SOURCE TESTING 125
C. Don Paulsell

VALIDATION OF CONTINUOUS STACK MONITORING DATA 131
Joseph E. McCarley

SCREENING CHECKS USED BY THE NATIONAL CLIMATIC CENTER 135
William E. Klint

DATA VALIDATION FOR UPPER AIR SOUNDING DATA AND EMISSION INVENTORY DATA 199
J. H. Novak

VALIDATION OF BIOMEDICAL DATA THROUGH AN ON-LINE COMPUTER SYSTEM 209
Larry D. Claxton

REGIONAL VALIDATION OF STATE AND LOCAL AIR POLLUTION DATA 219
Thomas H. Rose

DATA VALIDATION FOR THE LOS ANGELES CATALYST STUDY (LACS) 223
Charles E. Rodes

VALIDATION TECHNIQUES USED IN CONTINUOUS AIR MONITORING 237
Marvin B. Hertz

USE OF PRECISION AND ACCURACY ESTIMATES FOR VALIDATION OF DATA 247
David T. Mage

VALIDATION SYSTEM USED IN THE ST. LOUIS REGIONAL AIR MONITORING STUDY (RAMS) 265
Robert B. Jurgens

NAMES AND ADDRESSES:

PROGRAM 295

EPA/RTP INTERLABORATORY QUALITY ASSURANCE COORDINATING COMMITTEE 305

DATA VALIDATION CONFERENCE, SPEAKERS 306

DATA VALIDATION CONFERENCE, ATTENDEES 308

ACKNOWLEDGMENTS

The cooperation of all participants in the conference is gratefully acknowledged. Particular appreciation is due to the participants who prepared written copy of their presentations for post-documentation of the conference.

SECTION I

INTRODUCTION

A conference on data validation was held on November 4, 1977, at Research Triangle Park, North Carolina. The conference was sponsored, organized, and hosted by the Environmental Research Center - Research Triangle Park (ERC-RTP) Interlaboratory Coordinating Committee. Conference participants represented (a) EPA's RTP Research Laboratories and Program Offices, (b) an EPA Regional Office, (c) EPA Contractors, and (d) the National Climatic Center, Asheville, North Carolina.

Welcoming remarks were made by Dr. John K. Burchard, Director of the Industrial Engineering and Research Laboratory, and Senior ORD official at RTP. Each of the speakers presented their current practices for data validation. The conference provided an opportunity for a free exchange of viewpoints and techniques and was intended to enhance the state-of-the-art of Data Validation.

EXECUTIVE SUMMARY AND RECOMMENDATIONS

by

Seymour Hochheiser

and

Raymond C. Rhodes

Environmental Monitoring and Support Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

SECTION II

EXECUTIVE SUMMARY AND RECOMMENDATIONS

The nature and scope of data validation activities varies considerably among those involved with the collection, analysis, review, and use of environmental data for research and monitoring purposes. The objective of the conference was to review and discuss the various current practices of data validation, and to provide a forum for the exchange of information. It was intended that as a result of this conference, the function of data validation might become more specifically defined and more uniformly and widely implemented.

Following is a general review of the contents of the papers presented at the conference.

DEFINITIONS AND SCOPE OF DATA VALIDATION

The conference authors implied differing definitions of data validation and used various words relating to data validation activities. The scope of activities extended from that of simple checks for data transfer errors to that of a total quality assurance program. Words used in relation to data validation activities included: editing, screening, verification checking, auditing, and qualification.

TYPES OF DATA INVOLVED

The type of data involved in most of the papers was ambient air pollution concentrations. However, other types of data -- including epidemiological, meteorological, stationary source emission, mobile source emission, and in vitro and in vivo bioassays -- were discussed.

Data were validated in a variety of forms, including strip charts, handwritten forms, computer printouts, magnetic tape, and optical sensing records.

In some cases the activities of data validation, as indicated by the authors, were performed by those producing the data; in other cases, by independent reviewers; and in still other cases, by the users.

SIZE OF DATA BLOCKS CONSIDERED

In most cases, the data being validated were reviewed in definable blocks, varying from one day to one year. In one case, a real-time computerized system was used, and in another, the results of single tests were reviewed individually. The number of data values considered as a block, or group of data, varied from only three results for stack sampling tests to over 30,000 for pibals (meteorological balloons).

TYPES OF TECHNIQUES EMPLOYED

Both manual and computer techniques were used, usually depending upon the amount of data involved. Some systems employed both manual and computer methods. Only a few of the papers included graphical techniques for reviewing data. In less than half of the systems described, specific checks were made of the identification (or coding) of the data. In about half of the data validation systems, statistical techniques were used. Some of the techniques included were Dixon outlier tests, Shewhart control chart limits, exponential distributions, and asymptotic singular decomposition. In several instances, statistical sampling plans were utilized to select specific data sets for checking.

WRITTEN DATA VALIDATION PROCEDURES

In only one case were detailed procedures written describing the data validation activities and criteria.

FLAGGING AND REJECTION OF DATA

In most cases the questionable results of the data validation process were flagged, i.e., identified for more detailed evaluation and/or identified as questionable values in the data records. In most cases, as a result of data validation, questionable data were invalidated (rejected) or were corrected as a result of further investigation.

RECOMMENDATIONS

Although no official conference recommendations were made, the following recommendations were generally expressed:

1. The functions and scope of data validation should be more specifically defined.

2. Data validation techniques should be presented and summarized in some logical manner.

3. Data validation systems should be recommended or specified for use in certain situations.

The above recommendations could be pursued in several possible ways. One would be for a task group to be formed to develop standardized nomenclature, to summarize in a systematic way various activities and techniques of data validation, and to recommend data validation systems for specific situations. The above tasks could also be performed by a knowledgeable contractor.

SUMMARY

It is evident that the current practices of data validation vary widely in nature and scope. The conference provided an excellent opportunity for an open exchange of information concerning data validation practices and should result in a broader utilization of data validation techniques. In addition, the conference discussions and these proceedings should promote a greater awareness of the need to develop a more organized and unified approach to this important element of quality assurance for environmental data.

WHAT IS DATA VALIDATION?

by

Raymond C. Rhodes

Environmental Monitoring and Support Laboratory
U.S. Environmental and Protection Agency
Research Triangle Park, North Carolina 27711

WHAT IS DATA VALIDATION?

R.C.Rhodes

Just what is data validation?

Many of us are involved in activities which, we feel, constitute data validation, or at least, a part of a data validation process. My first encounter with the term "data validation" in connection with air pollution monitoring data occurred about five years ago. Since then my concepts of the function and scope of data validation have expanded considerably, and in fact they are still changing.

I'm sure that each person attending this conference has his or her own concept--probably different from anyone else's--of data validation. Whatever these concepts are, we're here to exchange our ideas, thoughts, and techniques on the subject. I feel sure that each of us will learn something new and useful for our own particular area of application.

Before we hear the other speakers, let's think a little bit about this subject of "data validation." Webster defines "validation" as follows:

VALIDATION

-- THE ACT OR PROCESS OF VALIDATING*

That doesn't help us very much, does it? So we might look at the definition of the word "valid" itself.

VALID

-- HAVING LEGAL EFFICACY OR FORCE

-- SUPPORTED BY OBJECTIVE TRUTH

---

*The capitalized items in this paper were used as visual aids for the presentation.

This definition is getting a little closer to our desired meaning in the data validation sense. The word "valid" does have some connotation of a "stamp of approval," indicating that things are "right."

In the "Quality Assurance Handbook for Air Pollution Measurement Systems," EPA 600/9-76-005, the following definition is given:

DATA VALIDATION

-- THE PROCESS WHEREBY DATA ARE FILTERED AND ACCEPTED OR REJECTED BASED ON A SET OF CRITERIA

There is a short section on "data validation" in the Handbook, which you may be interested in reading. My own definition, which I use in the "Data Validation" lecture of Air Pollution Training Institute (APTI) Course 470, "Quality Assurance for Air Pollution Measurement Systems," is somewhat more detailed:

DATA VALIDATION

-- A SYSTEMATIC PROCEDURE OF REVIEWING A BODY OF DATA AGAINST A SET OF CRITERIA TO PROVIDE ASSURANCE OF ITS VALIDITY PRIOR TO ITS INTENDED USE

The above definition says, in other words, that "a body of data" is reviewed according to some previously defined plan in a rather comprehensive and extensive way using all available expertise and knowledge at hand to assure that the data are technically consistent, correctly identified, and contain no obvious errors before the data are used.

Following are a number of terms which seem to involve functions or activities related to data validation.

RELATED TERMS

-- DATA EDITING
-- DATA SCREENING
-- DATA AUDITING
-- DATA VERIFICATION
-- DATA EVALUATION
-- DATA QUALIFICATION
-- DATA QUALITY ASSESSMENT

During the remaining presentations today, you will hear further references or usages of some of these terms. Since some of these terms are used interchangeably, I believe we need more specific definitions for each of the above to better understand how each one is, or is not, involved in data validation. As I define data validation and the above terms, I would include data editing, data screening, data auditing and data verification as part of data validation. However, according to my definitions, data evaluation, data qualification and data assessment are not parts of the data validation process.

Before considering some of the aspects of data validation, let us consider the obvious need for data validation and its relation to quality assurance. EPA and other organizations need good data from which to make good decisions. This truism applies equally well to research studies as well as to monitoring programs although data validation is not usually considered as a separate activity in research efforts.

GOOD DATA ⟶ GOOD DECISIONS

-- RESEARCH STUDIES
-- MONITORING PROGRAMS

Since data validation is concerned with an assurance of having obtained good data, one might think that data validation includes everything that is done to get valid, or good, data. But that is the concern of quality assurance. Whereas quality assurance is concerned with all activities which may affect data quality, the activities of data validation involve an after-the-fact review of the data, along with related information, to assure that valid data have, in fact, been obtained. As such, data validation is considered as only one element of quality assurance.

DATA VALIDATION

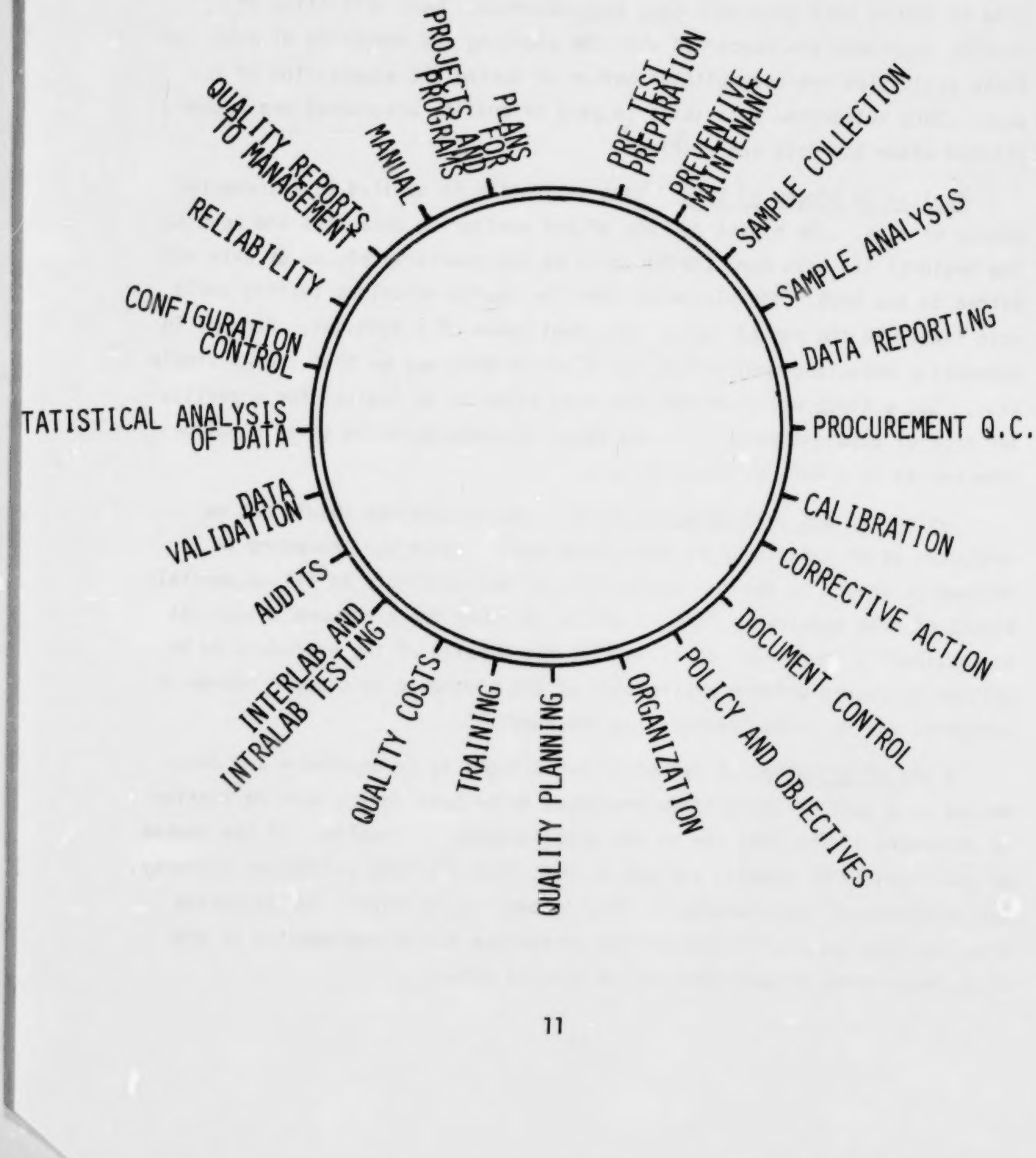
IS

AN ELEMENT OF
QUALITY ASSURANCE

In the APTI Course 470, data validation is one of 23 elements of quality assurance as shown by the following "Q.A. Wheel" in the Quality Assurance Handbook.

QUALITY ASSURANCE ELEMENTS AND RESPONSIBILITIES

(THE QUALITY ASSURANCE WHEEL)


PLANS FOR PROJECTS AND PROGRAMS
PRE TEST PREPARATION
PREVENTIVE MAINTENANCE
SAMPLE COLLECTION
SAMPLE ANALYSIS
DATA REPORTING
PROCUREMENT Q.C.
CALIBRATION
CORRECTIVE ACTION
DOCUMENT CONTROL
POLICY AND OBJECTIVES
ORGANIZATION
QUALITY PLANNING
TRAINING
QUALITY COSTS
INTERLAB AND INTRALAB TESTING
AUDITS
DATA VALIDATION
TATISTICAL ANALYSIS OF DATA
CONFIGURATION CONTROL
RELIABILITY
QUALITY REPORTS TO MANAGEMENT
MANUAL

What are some of the attributes of a data validation system? With no intent to restrict the other speakers concerning their views of data validation, following are some key features, in my opinion, of a data validation system.

After-the-Fact Review. Data validation is an after-the-fact review of data to assure that good data have been obtained. Many activities of quality assurance are concerned with the planning and acquiring of data, but these activities are accomplished before or during the acquisition of the data. Data validation activities (a part of quality assurance) are accomplished after the data are obtained.

Applied to Blocks of Data. Data validation is applied to incremental blocks of data. The blocks in case of air monitoring data that are sent to the National Air Data Bank (NADB) could be the quarterly blocks of data submitted to the NADB. The blocks of data for source emissions testing would most likely be the run of three individual tests of a test set. Perhaps in automotive emissions monitoring, the block of data may be that from a single test. So, a block would depend upon what seems to be logical for a particular type of data-gathering. In any case, the data would be given a validation review as a defined block of data.

Systematic and Uniform Application. Data validation should not be conducted on an occasional or spot-check basis. Once the procedure is defined it should be applied systematically and uniformly to all sequential blocks of data acquired. This is not to say that the procedure should not be continually improved. It is helpful for details of the procedure to be written to assure uniform application of the procedure in case of change of personnel and to avoid "reinventing the wheel."

A Set of Criteria. A set of criteria ought to be developed and documented as a part of the written procedure to be used during data validation to determine if the data are valid, questionable, or invalid. If the causes of questionable or invalid are not evident from the data validation activity, the detection of questionable or invalid data could trigger investigation into possible cause with appropriate corrective action implemented to preclude recurrence of questionable, or invalid data.

Checks for Internal Consistency. Data validation might include checks for internal consistency, such as relationships among pollutants, or relationships between pollutants and meteorology.

Checks for Temporal and Spatial Continuity. Data validation might include checks for continuity with respect to time, as might be evaluated by having a chronological plot of the data, to look for discontinuities, spikes, gaps, etc. The data may also have some spatial continuity if the data are from a network within some relatively small region, such as a local air monitoring network.

Checks for Proper Identification. To be useful, data must be properly identified. Improperly identified data may well be considered "no data." Although identification may seem to be a trivial thing, the Regions, for example, have difficulties with such improper identifications as (a) one state reporting data identified for another state, (b) data for October 35, and (c) duplicate data from one site and none from another. For medical history questionnaires of health effects studies, checks may be made to make sure that children are not older than their mothers!

Checks for Transmittal Errors. For paperwork systems, simple checks may be made to assure that the data have not been incorrectly transferred from one paper to another. With more sophisticated electronic and computer data handling and with telemetry of data, checks could be made to assure that the data have not been changed in the process.

Flagged or Rejected Data. A data validation system might include a scheme for flagging questionable data and may make provision for outright rejection of data for use. It may be desirable, however, to retain such data in the data system with proper indication of its status.

In summary, some of the aspects I consider as parts of a Data Validation System are as indicated below:

DATA VALIDATION

-- AFTER-THE-FACT REVIEW
-- APPLIED TO BLOCKS OF DATA
-- SYSTEMATICALLY AND UNIFORMLY APPLIED

- A SET OF CRITERIA
  - CHECKS FOR INTERNAL CONSISTENCY
  - CHECKS FOR TEMPORAL AND SPATIAL CONTINUITY
  - CHECKS FOR PROPER IDENTIFICATION
  - CHECKS FOR TRANSMITTAL ERRORS
- DATA FLAGGED OR REJECTED

Techniques of Data Validation. Obviously, because the methods of data gathering are so varied, the particular techniques that are to be used for data validation for a particular program will depend upon many things. Following are mentioned a few of the factors which need to be considered. The nature of the response output, that is, whether you get a response on a strip chart recorder, or whether it is generated on paper tape, magnetic tape, or is fed directly into a computer will determine the technique of data validation. The techniques will depend upon the method of data reduction, i.e., whether it is a manual-type method or a computer system. The form of the data transmittal, i.e., whether data are transmitted by some handwritten form, typewritten form, computer printout, or magnetic tape will determine the types of data checks to use. The techniques will also depend upon the amount of data. As we get involved with larger studies and larger blocks of data involved, such as NADB, different techniques must be used from those utilized for small sets of data. The techniques will depend upon the type and amount of ancillary (related) data that can be used for evaluation, comparison, or for correlation purposes. Techniques will depend upon what computing capability is available for use. The extent of available plotting capability is an important consideration, particularly for large blocks of data. Personally, I would like to see graphical presentations used in data validation. Much more can be learned by graphical representation that would be very difficult--almost impossible--to learn from visual review of large masses of data. Finally, the nature and extent of data validation techniques would depend on the intended use of the data. Different criteria may be used for validating data from which long term trends are estimated as compared to data for three-hour peak values, for example. To summarize:

TECHNIQUES WILL DEPEND ON

- -- NATURE OF RESPONSE
- -- METHOD OF DATA REDUCTION
- -- FORM OF DATA TRANSMITTAL
- -- AMOUNT OF DATA
- -- AMOUNT OF ANCILLARY DATA
- -- COMPUTING CAPABILITY
- -- PLOTTING CAPABILITY
- -- USE OF THE DATA

Lastly, there are two key principles of data validation that I want to mention. First, data validation ought to occur as close in time and location as possible to the originating location of the data. If questionable values are discovered, and corrective actions need to be made to the system, they must be made in a very timely and effective manner. For example, NADB may be validating data for as much as two years after the initial generation of the data. That is much too late to get effective corrective action at the local level. Therefore, data validation techniques should be located as closely as possible to the source of the data. Second, where possible, the persons having data validation responsibilities should not be the persons directly responsible for acquiring the data. Ideally, the person or persons responsible for data validation should be independent of the data acquisition activities and should be the most knowledgeable and experienced technical individual available to perform the function.

Thus,

DATA VALIDATION SHOULD BE

- -- CLOSE TO THE ORIGINATION OF THE DATA
- -- INDEPENDENT

Perhaps I have raised a number of questions in your mind concerning the subject of data validation. Hopefully, the other speakers will answer some of these questions, or will raise further questions, and will promote beneficial discussions and interchange of ideas and techniques of data validation.

THE SHEWHART CONTROL CHART TEST FOR
SCREENING 24-HOUR AIR POLLUTION MEASUREMENTS

by

William F. Hunt

Office of Air Quality Planning and Standards
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

THE SHEWHART CONTROL CHART TEST FOR SCREENING 24-HOUR AIR POLLUTION MEASUREMENTS

W.F. Hunt

INTRODUCTION

A quality control program is being developed for the U.S. Environmental Protection Agency's (EPA) National Aerometric Data Bank (NADB). The initial phases of the work were reported previously.(1,2) The purpose of the program is to develop and apply quality control tests to check ambient air quality data for anomalies, such as transcription and keypunch errors, as well as to detect erroneous data resulting from the periodic malfunctioning of air monitoring instruments. For the sake of completeness, it is worth reviewing some aspects involved in the collection and uses of air quality data. To begin with, air quality data are primarily collected to measure the success of emission control plans in achieving the National Ambient Air Quality Standards.

National Ambient Air Quality Standards (NAAQS)(3) have been established by EPA for five pollutants: total suspended particulate (TSP), sulfur dioxide (SO_2), carbon monoxide (CO), photochemical oxidants (O_x), and nitrogen dioxide (NO_2). These standards are intended to protect both human health and welfare. They may be stated as annual means or as upper limit values that may not be exceeded more than once per year. Although different averaging times are used for various standards, this paper is primarily concerned with the examination of 24-hour average values for TSP, SO_2, and NO_2 concentrations. While only TSP and SO_2 standards are in terms of 24-hour averages, all three pollutants have standards expressed in terms of annual averages. Because of the importance that is attached to violations of the NAAQS, a quality control program to ensure the validity of the measurement of both short- and long-term concentrations is extremely important.

The application of the Dixon Ratio Test(4) and Shewhart Control Chart Test(5) to measured levels of three major pollutants--TSP, SO_2, and NO_2--is examined. The tests apply to data from monitoring instruments which generate one measurement per 24-hour period and are operated on a systematic sampling schedule of approximately once every 6 days. In the cases of SO_2 and NO_2, there are also continuous monitoring instruments, which monitor the pollutants constantly; but our discussion here is concerned only with 24-hour data. The application of the tests results is flagged data which need to be verified as either valid or invalid.

These statistical tests are presently being applied to data collected in EPA's Region V. Region V encompasses the states of Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin. In terms of population, it is the largest of EPA's regions, and there is extensive monitoring of the above pollutants. The purpose of the Region V evaluation is to determine whether the data flagged by the tests are valid or invalid and to identify, if possible, the source of the error.

This paper will discuss the flow of data from the state and local government; the data-editing process; the basic characteristics of the data; and the application and evaluation of the two tests; it will conclude with our recommendations.

DATA FLOW

Most ambient air quality data are collected by state and local air pollution control agencies and are forwarded via EPA's Regional Offices to the NADB. A considerable amount of data is forwarded. For example, the minimum legal requirements for air pollution monitoring across the nation will result in the annual submittal of over 20 million air quality measurements to the NADB. The data are sent quarterly in a standard

Copyright ©1977 American Society For Quality Control, Inc. Reprinted by permission.

format(6) that specifies the site location; the year, month, and day of sampling; and the measurement itself (24-hour or 1-hour value) in micrograms or milligrams per cubic meter ($\mu g/m^3$ or mg/m^3) or parts per million (ppm). A corresponding site file contains descriptive information on the sampling-site environment. EPA edits the submitted data, checking for consistency with acceptable monitoring methods, and other identifying parameters. In the data-editing program, air quality data with extremely high values are flagged. Data that do not pass these checks or that have values exceeding certain predetermined limits are returned to the originating agency via the Regional Office for correction and resubmittal.

As might be expected with data sets this large, there are still anomalous measurements that slip through the existing editing and validation procedures. Therefore, there is a need for a simple cost-effective statistical test that can be applied to the air quality data by which to detect, primarily, obvious transcription, keypunch, and measurement errors. Statistical tests do not eliminate, however, the need for more intensive quality assurance at the local level. For example, inadequate calibration procedures or similar problems that result in measurement bias will not be detected by our statistical procedures, which are intended primarily for macroanalysis.

BASIC CHARACTERISTICS OF TSP, SO_2, AND NO_2 DATA

Basic characteristics of the TSP, SO_2, and NO_2 data were considered in selecting the quality control tests being used. To begin with, the tests were applied to data which were obtained from monitoring instruments that generate one measurement per 24-hour period.(7) For such monitoring methods, EPA recommends that a systematic sampling procedure of once every 6 days, or 61 samples per year, be used at a minimum to collect the data.(8) Such a sampling procedure generates data, which for our purposes, may be considered as approximately independent.

In examining the distributional properties of the data, past research has shown that ambient TSP concentrations are approximately lognormally distributed.(9,10) This is sometimes true for SO_2 and NO_2, also, but is not always the case. Current work suggests that these pollutants may follow an exponential or Weibull distribution.(11)

In selecting the quality control tests, the averaging times which correspond to the NAAQS are important. The values of interest are the peak concentrations (24-hour average measurements) for TSP and SO_2, and the annual means for TSP, SO_2, and NO_2.

The final data characteristic of importance is the seasonality of the pollutants. As an example, in some areas of the country, TSP and SO_2 measurements are highest in the winter months and lowest in the summer months. Therefore, the factor of seasonality had to be considered in the selection of the quality control test to minimize this as a possible source of error.

THE QUALITY CONTROL TESTS

Two quality control tests are presently being applied and the results of the application evaluated, the Dixon Ratio Test(4) and the Shewhart Control Chart Test.(5) The output of the quality control tests is a listing of the suspicious data, including the site and the time of occurrence. The tests are discussed below.

Dixon Ratio Test

The use of the Dixon Ratio Test was discussed in an earlier paper.(2) The test was applied to TSP quarterly data and was found to work reasonably well in detecting a single anomalous value. Problems occurred when there were multiple transcription errors within a quarter, such as the miscoding of an entire month of data. This problem was corrected when the test was applied to monthly averages.

As part of the evaluation of quality control of Region V data, the Dixon Test was applied to all 1974 monthly averages of TSP, SO_2, and NO_2 on a site-by-site basis to examine the data for possible multiple transcription, keypunch, or measurement errors occurring within a month. By applying the test to the monthly averages, the assumption of normality can be satisfied, although the monthly averages are not entirely independent because of the seasonality in the data. This must be considered in examining the flagged data.

The Dixon Ratio Test requires that the monthly averages be ordered in increasing levels of magnitude. The test basically constructs an "r" ratio that compares the distance of the maximum (minimum) observation from its neighbors with the range of all but one or two of the observations. Let us assume that Y_i equals the i^{th} order pollutant monthly average, where Y_N is the highest monthly average and N equals the number of months within the year for which there are data. The test procedure is as follows:

1. Choose α, the probability (risk) of rejecting an observation that really belongs in the group.
2. Order the monthly averages from Y_1 through Y_N, where Y_N is the highest value.
3. If $3 \le N \le 7$, compute $r_{10} = (Y_N - Y_{N-1})/(Y_N - Y_1)$;

   $8 \le N \le 10$, compute $r_{11} = (Y_N - Y_{N-1})/(Y_N - Y_2)$;

   $11 \le N \le 12$, compute $r_{21} = (Y_N - Y_{N-2})/(Y_N - Y_2)$;

   where Y_N is the highest value.
4. Look up $r_{1-\alpha}$ for r_{ij} from a table of critical values.[4]
5. If r_{ij} is greater than $r_{1-\alpha}$, print out a list showing the suspect monthly averages, the remaining monthly averages, and the site location.

Reproduced from best available copy.

The Shewhart Control Chart Test(5) can be used to examine both shifts in monthly averages, as well as shifts in the monthly range. From the former it can detect possible multiple errors and from the latter, single anomalous values. In this test the data can be divided up into what Shewhart called rational subgroups.(12) In a manufacturing process the subgroups would most likely relate to the order of production. Ambient air quality measurements can be viewed in the same way because they are collected by a monitoring instrument over time. A month of data was selected as the rational subgroup because the air quality data are recorded by the state and local agencies on a monthly basis in a standard format.(6) The monthly subgroup generally consists of five measurements based on EPA's recommended sampling schedule(8) of 61 observations per year, which also is the common subgroup size found in industrial use.(5) Using a subgroup size of five, it can be assumed that the distribution of the monthly means is nearly normal, even though the samples are taken from a nonnormal universe.

The test was applied to the 1974 Region V data on a moving 4-month basis: that is, the averages and range of values in the month in question were compared with the overall averages of the three previous monthly averages and monthly ranges. The moving 4-month comparison was used to minimize the effect of the seasonality of the pollutants. The formulas for calculating the trial limits are as follows:

For the monthly range: $UCL_R = D_4\bar{R}$, and

$LCL_R = D_3\bar{R}$.

For the monthly means: $UCL_{\bar{x}} = \bar{\bar{x}} + A_2\bar{R}$, and

$LCL_{\bar{x}} = \bar{\bar{x}} - A_2\bar{R}$,

where R = the monthly range; $\bar{R}$ = the average of the three previous monthly ranges; $\bar{x}$ = the monthly average in question; $\bar{\bar{x}}$ = the average of the three previous monthly averages, and D_3, D_4, and A_2 are factors for determining from $\bar{R}$ the 3-sigma control limits for $\bar{x}$ and R. (See Table C on page 562, reference number 5.)

RESULTS OF APPLICATION OF QUALITY CONTROL TESTS

During 1974, TSP, SO_2, and NO_2 were being monitored in Region V at 855, 366, and 303 sites, respectively. Both the Dixon and Shewhart Tests were applied to all 1974 TSP, SO_2, and NO_2 data from Region V. Still in progress, an extensive effort is being made on the part of EPA personnel in Region V, in conjunction with state air pollution control officials, to evaluate the air quality data flagged by both the Dixon and Shewhart Tests. As an initial phase of this evaluation, examination was made of those data in which the flagged monthly mean or range exceeded one of the pollutant-specific NAAQS. For TSP and SO_2, appropriate cutoffs were thought to be 260 $\mu g/m^3$ and 365 $\mu g/m^3$, which are their respective primary short-term 24-hour standards. In the case of NO_2, the annual primary NAAQS of 100 $\mu g/m^3$ was used because NO_2 has no short-term primary standard. Although their choice was somewhat arbitrary, the NAAQS were used as cutoffs because their violation results in reexamination of the overall adequacy of local air pollution control measures in effect. Thus, high values must be verified because they can result in significant impact on the original control strategy designed to achieve the NAAQS.

Table 1 indicates the number of Region V sites reporting TSP, SO_2, and NO_2 data which were flagged by the Dixon Test, by the Shewhart Control Test, and by both tests. As would be expected, there are more sites flagged by the Shewhart Control Test as having anomalous data than the Dixon Test, because it looks at both shifts in the monthly mean and range while the Dixon Test examines only the monthly means. The preliminary evaluation of the flagged sites is also given as the number of flagged sites which were found to have one or more erroneous 24-hour measurements.

TABLE 1. Comparison of Dixon Ratio and Shewhart Control Chart Tests as Applied to Sites in Region V Monitoring TSP, SO_2, and NO_2 in 1974

	Pollutant		
	TSP	SO_2	NO_2
High value in question[a] ($\mu g/m^3$)	≥ 260	≥ 365	≥ 100
Total sites, no.	855	366	302
Dixon test			
Flagged sites, no.	35	1	25
Flagged sites, no. with errors	31	1	11
Shewhart test			
Flagged sites, no.	38	4	36
Flagged sites, no. with errors	31	3	16
Both tests			
Flagged sites, no.	32	1	19
Flagged sites, no. with errors	31	1	10

[a] The High value in question is the monthly mean in the case of the Dixon Test and the monthly mean or range in the Shewhart Control Chart Test. The National Ambient Air Quality Standards (NAAQS) were used as high value cutoffs: 260 $\mu g/m^3$ and 365 $\mu g/m^3$ are the 24-hour primary NAAQS for the TSP and SO_2, respectively, while 100 $\mu g/m^3$ is the annual primary NAAQS for NO_2.

Of the 855 sites in Region V measuring TSP in 1974, 35 were flagged by the Dixon Test, 38 by the Shewhart Control Test, and 32 by both tests. The flagged sites reported at least one monthly mean and/or range equal to or greater than 260 $\mu g/m^3$. The preliminary evaluation indicates that data from 31 sites, which were flagged by both tests, were found to have multiple transcription or keypunch errors. In the case of SO_2, 1 of the 366 sites was flagged by the Dixon Test, 4 by the Shewhart Test, and 1 by both tests. The monthly mean and ranges in question were equal to or greater than 365 $\mu g/m^3$. Data from the site flagged by both tests were found to have multiple transcription errors, while data from the remaining two sites flagged by the Shewhart Test had single transcription errors. Finally, of the 302 sites measuring NO_2, 25 were flagged by the Dixon Test, 36 by the Shewhart Test, and 19 by both tests. The monthly means and ranges in question equalled or exceeded 100 $\mu g/m^3$. Transcription and keypunch errors were found at 11 of the sites flagged by the Dixon Test, 16 of the sites flagged by Shewhart Test, and 10 of the sites flagged by both.

An example of a site flagged by both tests was one that measured TSP for 11 months in 1974. The monthly mean ($\bar{X}$), ranges (R), and subgroup sizes (n) are indicated below by month:

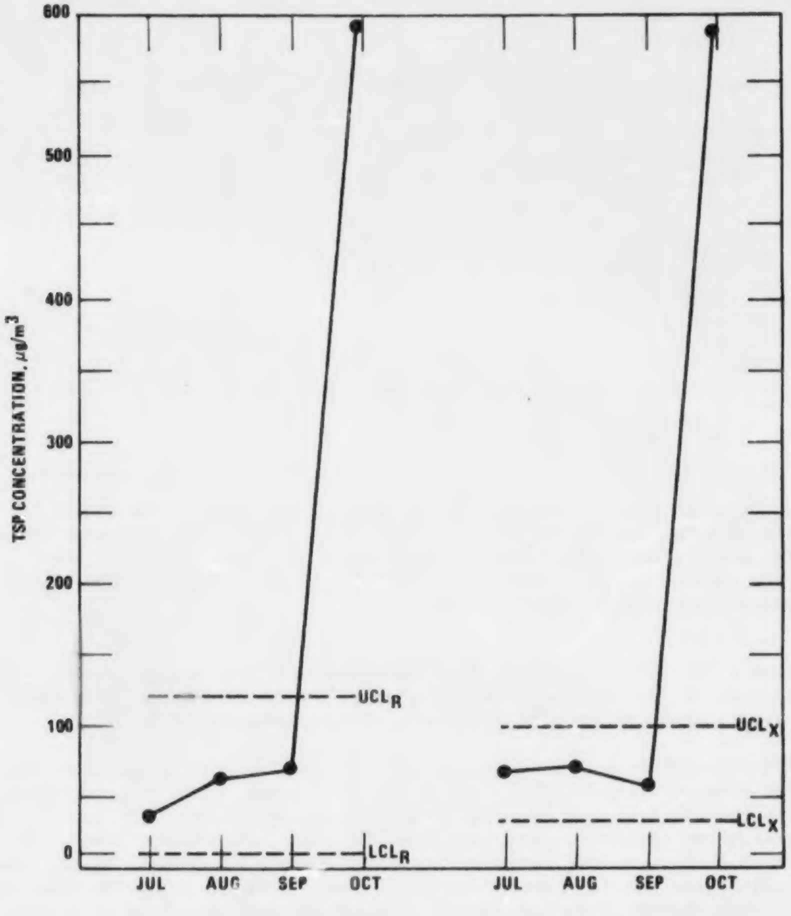
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
$\bar{x}$	-	67	60	56	70	67	66	73	59	591	82	41
R	-	74	25	71	44	102	37	64	68	595	68	30
n	0	4	5	5	5	3	5	5	5	5	3	4

The Dixon Ratio Test was applied to the entire year of data; the ratio of the largest monthly mean, 591, minus the third largest mean, 73, was compared with the difference of the largest mean and second smallest monthly mean, 56. The test statistic is

$$r_{21} = \frac{591-73}{591-56} = 0.968$$

which is significant at the $\alpha = 0.005$ level.

The Shewhart Control Chart Test was applied on a moving 4-month basis. When the monthly average and range for October became the values in question, they were compared with the overall averages of the July, August, and September averages and ranges. The test results are shown in Figure 1 for both the monthly mean and range. In both cases the air quality data are "out of control" for the month of October, with both the October average and range way above their respective upper control limits. The problem was later identified as multiple transcription errors in which all numbers in the month of October were off by a factor of 10.



a. R CHART FOR MONTHLY RANGE

b. X CHART FOR MONTHLY MEAN

Figure 1. Example of Shewhart Control Chart Test applied to data with multiple transcription errors in month of October.

CONCLUSION

From the initial results of the Region V evaluation, it appears that both the Dixon and Shewhart work well on the TSP, SO_2, and NO_2 data and are in reasonably good agreement. Ideally, both tests should be used in the screening process. However, if an air pollution control agency wanted to employ only one test, the Shewhart Control Chart Test would be preferable, because it has the advantage that it can simultaneously examine shifts in both the monthly mean and range and can be presented graphically. Further, in the case of SO_2 and NO_2, the Shewhart Test flagged sites with a single transcription or keypunch error--identified by shifts in the range--which were not identified by the Dixon Test.

The second phase of the Region V evaluation will cover those sites whose highest measured value did not exceed one of the pollutant-specific NAAQS. This phase will be examined in a later paper, along with the development of quality control tests for data generated by the continuous monitoring methods.(7)

ACKNOWLEDGMENTS

The authors wish to express their appreciation to the state air pollution control agencies in Region V for their help in the evaluation of the tests, to Mrs. Ann Rogers and Mrs. Aline Rolaff for providing the computer programming support, to Mrs. Joan Bivins, Miss Hazel Browning, and Mr. Willie Tigs for their clerical support, and to Dr. Thomas Curran for his many helpful comments on earlier drafts of the paper.

REFERENCES

1. Hunt, W. F., Jr., and T. C. Curran. An Application of Statistical Quality Control Procedures to Determine Progress in Achieving the 1975 National Ambient Air Quality Standards. Transactions of the 28th Annual ASQC Conference, Boston, Massachusetts, May 1974.

2. Hunt, W. F., Jr., T. C. Curran, N. H. Frank, and R. B. Faoro. Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air. Transactions of the Joint European Organization for Quality Control/International Academy for Quality Conference, Venice Lido, Italy, September 1975.

3. Title 40 - Protection of Environment. National Primary and Secondary Ambient Air Quality Standards. Federal Register. 36(84):8186-8201, April 30, 1971.

4. Dixon, W. J. Processing Data for Outliers. Biometrics. 9:75, 1953.

5. Grant, E. L. Statistical Quality Control. New York, McGraw Hill Book Co. p. 122-128. 1964.

6. SAROAD Users Manual. U. S. Environmental Protection Agency, Research Triangle Park, N.C. Publication No. APTD-0663. July 1971.

7. Hoffman, A. J., T. C. Curran, T. B. McMullen, W. M. Cox, and W. F. Hunt, Jr. EPA's Role in Ambient Air Quality Monitoring. Science. 190(4211):243-248, October 1975.

8. Title 40 - Protection of Environment. Requirements for Preparation, Adoption, and Submittal of Implementation Plans. Federal Register. 36(158):15490, August 14, 1971.

9. Larsen, R. I. A Mathematical Model for Relating Air Quality Measurement to Air Quality Standards. U. S. Environmental Protection Agency, Research Triangle Park, N.C. Publication No. AP-89. 1971.

10. Hunt, W. F., Jr. The Precision Associated with the Sampling Frequency of Lognormally Distributed Air Pollutant Measurements. J. Air Poll. Control Assoc. 22(9): 687, 1972.

11. Curran, T. C. and N. H. Frank. Assessing the Validity of the Lognormal Model When Predicting Maximum Air Pollutant Concentrations. Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston, Massachusetts, 1975.

12. Shewhart, W. A. Economic Control of Quality of Manufactured Product. Princeton, D. Van Nostrand Company, Inc. 1931. p. 299.

DISTRIBUTION GAP TEST FOR HOURLY AIR POLLUTION DATA

by

Thomas C. Curran

Office of Air Quality Planning and Standards
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

DISTRIBUTION GAP TEST FOR HOURLY AIR POLLUTION DATA

T.C. Curran

Previous papers[1-3] have discussed techniques for screening air pollution data sets with particular attention given to 24-hour measurements. The present paper focuses upon the use of screening procedures for hourly ambient air quality measurements. As with any quality control procedure, it is useful to consider the nature and intended use of the data before discussing the screening technique.

Hourly air pollution data sets present some interesting practical problems when one considers the use of a screening procedure. The most obvious feature is the volume of data. For example, 24-hour air pollution measurements are usually obtained by every-sixth-day sampling resulting in approximately 60 samples per year. In contrast, hourly measurements are obtained from continuous monitors that operate every day and, therefore, may produce as many as 8,760 values per year. Thus, hourly data sets are commonly 100 times larger than those for daily measurements. The reason that the volume of data is important becomes apparent when the use of the data is examined. For the most part, air pollution data is collected to determine status with respect to certain legal standards, such as the National Ambient Air Quality Standards.[4] These standards specify upper limits for air pollution concentrations. Of particular interest for this paper are the standards for oxidants or carbon monoxide which indicate hourly values "not to be exceeded more than once per year."[4] In these situations it is the second highest value from a data set of 8,760 observations that becomes the decision-making value. Obviously, this places a premium on ensuring data quality.

From a practical viewpoint, maintaining a data bank for air pollution measurements involves the basic conflict of having to routinely process large volumes of data and yet at the same time ensure an almost zero defect level of data quality. Many sites monitor for several pollutants so that on the national level, thousands of sites are routinely submitting tens of thousands of data points each year. However, because of the nature of the standards, many users may only be interested in the two highest values at each site for each pollutant. It should be noted that two values from a data set of 8,760 observations constitutes 0.023 percent of the data. This means that the user's perception of data quality may be entirely different from the true data quality. For example, if only 0.05 percent of the data points were too high due to errors, this would still be sufficient to have the user complain that "the data are useless." On the other hand, if elaborate editing checks are introduced, the sheer volume of data may result in high costs or processing delays, and the user may now complain that the data are not sufficiently current for him to make timely decisions.

With this background in mind, it is apparent that an air quality data screening program must be able to process large volumes of data in an inexpensive fashion while flagging virtually every error. Also, because it is frequently difficult and time consuming to verify suspect data points, every flagged value should be a genuine error. Unfortunately, while these characteristics are obviously desirable, they are also almost impossible to attain. The approach presented here is primarily intended to eliminate the more glaring errors from these hourly data sets. The major emphasis is on screening the higher concentration values to check for general internal consistency within the data set.

RATIONALE FOR SCREENING PROCEDURE

In our initial development of a screening procedure for hourly data, a computer program was developed that checked for departures from typical patterns.[3] These typical patterns were selected on the basis of experience with various types of air pollution data. Basically, the values were flagged on a yes-no decision, and there was no probability statement associated with the rejected values. One stage in this development was

Copyright © 1977 American Society For Quality Control, Inc. Reprinted by permission.

to give sample data sets to experienced air pollution data analysts to see what values they would reject. There were two reasons for this step. The most obvious was to ensure that the computerized screening procedure was consistent with so-called expert judgment. However, another reason was the need for a test that would mimic the decisions made by an experienced analyst. The reason for this was an attempt to avoid a black-box approach where the screening procedure was viewed as a mysterious oracle delivering arbitrary decisions. The point here is that it can be quite time consuming for the data analyst to check flagged data points. Values that appear to be quite unlikely from a statistical viewpoint may actually be quite likely in the real world. For example, massive traffic jams do happen and may result in high carbon monoxide levels. Windstorms can mean high total suspended particulate levels. Sudden shifts in wind direction can mean that a monitor near a point source goes from a zero reading to almost full scale and back in a few hours. The high variability associated with peak air pollution values makes it almost impossible to develop a screening procedure that does not occasionally flag real values. But it seemed desirable to avoid the situation where an air pollution analyst would tire of repeatedly checking flagged values that turned out to be correct. Therefore, emphasis was given to developing a test that would flag values that an air pollution analyst would want to investigate. An effective way to accomplish this was to develop a test that would mimic experienced human judgment so that the analyst would understand why the value was flagged.

To a large degree the preliminary test on patterns was successful. Experienced analysts used the same basic approach of looking for unusual jump discontinuities between successive hourly values or departures from expected diurnal or seasonal patterns. However, there were two main deficiencies in this computerized procedure based upon departures from suspected patterns. One was the lack of a probabilistic framework. The second, and probably the more serious from a practical standpoint, was the need to vary the amount of allowable departure from site to site. The probabilistic framework could be provided by a time series model, and the parameters varied from site to site. However, it became apparent during the preliminary investigation that many of the outliers could be detected by a much simpler approach. In most cases, unusually high values could be detected by examining the frequency distribution of the hourly data for a given period of time, such as a month, quarter, or year. Suspect values would be associated with large gaps in the frequency distribution. The length of the gap and the number of values above the gap afforded a convenient means of detecting possible errors. With this simplification of the problem, it becomes possible to develop a probabilistic framework for the problem as discussed below.

PROBABILITY OF A GAP

In order to compute the probability of a gap in the empirical frequency distribution, it is necessary to assume some type of underlying distribution. Although this involves an oversimplification because it ignores dependency between successive hourly values, such approaches have traditionally been used with success in air pollution data analysis.[5] The lognormal distribution has customarily been used for this purpose. However, the exponential distribution has also been found to provide a reasonable approximation for the upper tail, or higher concentrations, of hourly air pollution data.[6] Because the higher concentration values were of primary interest and the exponential distribution is mathematically convenient, it was used as the underlying distribution. As with any measurements, although the approximating distribution is continuous, the air pollution values are discrete valued. For simplicity, they may be assumed to be integers because this involves merely a change of scale. A gap in the frequency distribution may then be described in terms of its length, the number of values above the gap, and at what concentration the gap begins. Therefore, if a monthly empirical frequency distribution of hourly values has n values greater than concentration c but no values between c, and c+k, this would be a gap of length k starting at c with n observations above the gap. To compute the probability of this event, consider the following:

Let X be an exponential random variable.

Then $\Pr(X \le c) = 1-e^{-\lambda(c-\theta)}$ where $\lambda>0$, $c \ge \theta$.

Thus, $\Pr(X>c) = e^{-\lambda(c-\theta)}$.

The probability that X is greater than c+k given that X is greater than c is

$$\Pr(X>c+k \mid X>c) = \frac{\Pr(X>c+k)}{\Pr(X>c)} = \frac{e^{-\lambda(c+k-\theta)}}{e^{-\lambda(c-\theta)}} = e^{-\lambda k}$$

Because X is distributed exponentially, this expression is independent of the concentration c.

Assuming independence, the probability that n values are greater than c+k given that these n values are greater than c is

$$(e^{-\lambda k})^n = e^{-n\lambda k} .$$

Thus, the probability of a gap of length k with n values above the gap is $e^{-n\lambda k}$. This probability then becomes the criteria for rejecting suspect data.

APPLICATION

A relatively simple FORTRAN program was written to process hourly data, compute the empirical frequency distribution, and examine any gaps. Because of the manner in which the data is routinely submitted to the U.S. Environmental Protection Agency's National Aerometric Data Bank, the program was written to check the data on a monthly basis (744 hourly values). The parameter λ obviously varies from one data set to another. For simplicity, λ was determined from the 50th and 95th percentiles of the data. This was computationally convenient and also emphasized the fit for the upper tail. Results to date in evaluating this test indicate that this approach is adequate.

Past experience has indicated that an occasional source of error is the miscoding of units so that an entire month of data would be internally consistent yet too high by some scale factor. To account for this, a second estimate of λ was computed using an assumed value for the 99.9th percentile, i.e., a value that historically should not be exceeded more than one time in a thousand.

RESULTS

In order to provide a realistic test of this screening procedure, actual data sets were used. One of particular interest involved carbon monoxide data that had been quickly key-punched and then manually edited for a specific study. This provided a preliminary and corrected version of the file. The preliminary file had known errors and the corrected file was presumably valid. The first test run on the preliminary file processed 21,362 hourly values from 40 monthly data sets. Eight of these monthly data sets were flagged. Hourly carbon monoxide values would be expected to mostly fall in the range of 0 to 50 ppm. In this first test, values of 900, 800, 700, and 500 were found resulting in gap lengths greater than 100 and associated probabilities of less than 1 in 10,000. These results are shown in Table 1. Of the eight flagged data sets,

TABLE 1. Rejected Site Months From Sample Data Set

Site	Month/year	Number of values	Maximum	2nd high	Gap length	Starting at	Number of values above	Probability
33	Oct. 1974	530	30	13	16	14	1	.0006
33	Nov. 1974	604	500	300	>100	15	3	<.0001
33	Dec. 1974	671	800	500	>100	41	4	<.0001
33	Jan. 1975	653	500	500	>100	20	2	<.0001
33	Feb. 1975	510	33	18	14	19	1	.0001
39	June 1975	707	900	700	>100	27	3	<.0001
901	July 1974	620	15	14	3	11	3	.0056
901	Aug. 1974	334	800	800	>100	11	5	.0001

seven had keypunch errors. The one remaining month was flagged on the basis of a gap of length 3 and the data appeared to be reasonable. This presented no difficulty for the analyst because the computer printout was sufficient to indicate that these data were in an intuitively acceptable range and probably did not warrant further investigation.

It took less than 30 seconds on EPA's UNIVAC 1110 to process these 21,362 hourly values, and the total cost was approximately $1. It should be noted that the program does several other editing checks so that this cost includes more than the screening procedure for gaps.

CONCLUSIONS

Using gaps in monthly frequency distributions appears to be a convenient means of screening hourly air pollution data sets for outliers. Results to date indicate that it satisfies the criteria of being easy and economical to implement while producing output that is intuitively understandable to an air pollution data analyst. The test successfully spots the more obvious errors. As expected, the initial results also suggest that these types of data sets do have a much lower error rate than the user perceives because of the emphasis on only the few highest values.

There are certain refinements that can be made in screening these type of data sets. Time series models and the use of associated data, such as meteorological variables, would be expected to increase sensitivity and possibly result in even better data quality. However, it remains to be seen if these more elaborate approaches are cost effective when processing vast quantities of data from locations throughout the nation.

As a final comment, it should be noted that once a value is flagged as a possible anomaly, it cannot be arbitrarily dropped from the data set. It must first be verified that the data point actually is incorrect. The fact that the data point is statistically unusual does not necessarily mean that it did not occur.

REFERENCES

1. Hunt, W. F., Jr., and T. C. Curran. An Application of Statistical Quality Control Procedures to Determine Progress in Achieving the 1975 National Ambient Air Quality Standards. Transactions of the 28th Annual ASQC Conference, Boston, Massachusetts, May 1974.

2. Hunt, W. F., Jr., T. C. Curran, N. H. Frank, and R. B. Faoro. Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air. Transactions of the Joint European Organization for Quality Control/International Academy for Quality Conference, Venice Lido, Italy, September 1975.

3. Hunt, W. F., Jr., R. B. Faoro, and S. K. Goranson. A Comparison of the Dixon Ratio Test and Shewhart Control Chart Test Applied to the National Aerometric Data Bank. Presented at the 30th Annual Conference of the American Society for Quality Control. Toronto, Ontario, Canada, June 1976.

4. Title 40 - Protection of Environment. National Primary and Secondary Ambient Air Quality Standards. Federal Register. 36:(84):8186-8201, April 30, 1971.

5. Larsen, R. I. A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards. U.S. Environmental Protection Agency, Research Triangle Park, N.C. Publication No. AP-89. 1971.

6. Curran, T. C. and N. H. Frank. Assessing the Validity of the Lognormal Model when Predicting Maximum Air Pollutant Concentrations. Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston, Massachusetts, 1975.

USE OF STATISTICAL SAMPLING IN VALIDATING HEALTH EFFECTS DATA

by

Carolyn P. Chamblee

Health Effects Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

USE OF STATISTICAL SAMPLING IN VALIDATING HEALTH EFFECTS DATA

Carolyn P. Chamblee
Statistics and Data Management Office
Health Effects Research Laboratory
Research Triangle Park, North Carolina 27711

ABSTRACT

A quality control plan has been adopted for large computer data files of health effects research studies. The Dodge-Romig acceptance sampling technique was selected. This procedure has the capability of guaranteeing within specific tolerance limits the agreement between the information on the computer files and the information on the original data documents. The method is easy to use and is adaptable to a wide range of files and to a varying quantity of documents. The type of plan chosen utilizes a file as a lot, a single document as a characteristic, a single sampling procedure, and a 2% Lot Tolerance Per Cent Defective (LTPD). Our experience with this acceptance sampling plan has been positive enough that we have extended its use to most of our current studies.

USE OF STATISTICAL SAMPLING IN VALIDATING HEALTH EFFECTS DATA

Carolyn P. Chamblee
Statistics and Data Management Office
Health Effects Research Laboratory
Research Triangle Park, North Carolina 27711

The Statistics and Data Management Office supplies statistical and data processing support to the Health Effects Research Laboratory (HERL) as required. One of the principal responsibilities of the Laboratory, as the name suggests, is to research and assess effects of air·pollution on the human health. One method HERL uses to carry out its responsibility is to conduct nationwide epidemiological research to establish the relationship between human health and community air quality. This research includes field studies that examine the health of population groups residing in communities exposed to definable air pollutants. Exposure-response relationships and injury thresholds are estimated and the studies document changes in health that accompany changes in environmental quality.

Questionnaires are designed for the field studies to allow more uniform collection of data. These questionnaires are usually designed as keypunch entry or optical scanning documents or a combination of both. As one would imagine for nationwide studies, the collection effort results in a large volume of information which is then processed through various steps and results in a computerized master file or files.

During the period from 1970 to 1975, HERL conducted a large number of epidemiological studies commonly referred to as CHESS or the Community Health and Environmental Surveillance System. There are approximately 83 of these studies covering five different types of studies over five areas. While successfully completing this intensive data collection effort during a period of in-house personnel limitations, computer conversion from an IBM 360/50 to a Univac 1110 and high contractor personnel turnover, HERL developed a backlog of raw data. Although these data were computerized and computer-edited, no one could make a definitive statement regarding the accuracy of the files versus the original source documents. The emphasis and importance of quality control procedures and our inability to qualify the data files led us to the point that a quality control program for these files had to be developed. I was assigned to develop the quality control plan.

The plan selected had to be able to guarantee that when properly followed the contents of the computer file reflected data reported on the forms within a small error tolerance. Also, each file must meet the error tolerance. That is to say, a statement that over the 83 files the error rate is less than the specified tolerance is not sufficient. Each individual file must satisfy the limit. Lastly, the quality control plan had to minimize the verification effort but must also be simple to use, easy to understand and adaptable over a wide range of data files and for a varying number of data forms.

A number of statistical and quality control references were reviewed before it was decided that the Dodge-Romig acceptance sampling technique was the most desirable. This approach is discussed in detail by Harold F. Dodge and Harry G. Romig (1). Their book is very easy to read and understand and offers a twelve step procedure for selecting a specific sampling plan. To explain how we decided on the plan we currently use, I will briefly describe the steps and discuss how we implemented Dodge-Romig.

1. Decide what characteristics to include. For example, a characteristic could be considered a variable or a data field, which could lead to distinctly different error rates. In our case, we considered all information on a single questionnaire form as a group so that one form equals one record.

2. Decide what is to constitute a lot. A lot is defined as a homogeneous material unit from a common source. In choosing the lot unit we balanced the fact that a small number of large lots can shorten inspection time against the additional difficulty of processing the rejected lots. In our case a lot equals one file.

3. Choose the type of protection. There are two types of protection: Lot Tolerance Per Cent Defective (LTPD) and Average Outgoing Quality Limit (AOQL). The AOQL applies to the average level of quality over all lots being inspected. It is appropriate for a continuing supply of a product. The LTPD applies to the quality level of each individual lot. We chose the LTPD type of protection.

4. Choose a suitable level of LTPD or AOQL. For LTPD choose the value of per cent defective you are willing to accept not more than 10 per cent of the time, that is, reject at least 90 per cent of the time. We balance the inspection costs against the consequences of accepting a file of bad quality. We considered rates in the range of 1% to 3% LTPD. We decided that 1% error rate was too costly and selected a rate of 2% LTPD.

5. Choose between single sampling and double sampling. For better economy in an overall inspection effort, double sampling is usually preferable. However, for minimum variation in the workload, single sampling should be used. We selected single sampling as a more straight forward and preferable method in our case.

6. Select the proper sampling table on the basis of the preceeding choices. We selected the Single Sampling Table for LTPD = 2 per cent (Figure 1 reproduced from reference 1).

7. Obtain an estimate of the Process Average Per Cent Defective (PA). Use previous data to obtain the PA. Even a rough estimate should be used if little prior data are available. A poor estimate will only decrease the economy of the plan but maintains the same LTPD protection. After some initial examination of HERL data, the column entitled "Process Average 0.61% to 0.80%" was used.

SINGLE
SAMPLING
2.0%
LTPD

Single Sampling Table for
Lot Tolerance Per Cent Defective (LTPD) = 2.0%

Lot Size	Process Average 0 to 0.02%			Process Average 0.03 to 0.20%			Process Average 0.21 to 0.40%			Process Average 0.41 to 0.60%			Process Average 0.61 to 0.80%			Process Average 0.81 to 1.00%		
	n	c	AOQL %	n	c	AOQL %	n	c	AOQL %	n	c	AOQL %	n	c	AOQL %	n	c	AOQL %
1-75	All	0	0	All	0	0	All	0	0	All	0	0	All	0	0	All	0	0
76-100	70	0	0.16	70	0	0.16	70	0	0.16	70	0	0.16	70	0	0.16	70	0	0.16
101-200	85	0	0.25	85	0	0.25	85	0	0.25	85	0	0.25	85	0	0.25	85	0	0.25
201-300	95	0	0.26	95	0	0.26	95	0	0.26	95	0	0.26	95	0	0.26	95	0	0.26
301-400	100	0	0.28	100	0	0.28	100	0	0.28	160	1	0.32	160	1	0.32	160	1	0.32
401-500	105	0	0.28	105	0	0.28	105	0	0.28	165	1	0.34	165	1	0.34	165	1	0.34
501-600	105	0	0.29	105	0	0.29	175	1	0.34	175	1	0.34	175	1	0.34	235	2	0.36
601-800	110	0	0.29	110	0	0.29	180	1	0.36	240	2	0.40	240	2	0.40	300	3	0.41
801-1000	115	0	0.28	115	0	0.28	185	1	0.37	245	2	0.42	305	3	0.44	305	3	0.44
1001-2000	115	0	0.30	190	1	0.40	255	2	0.47	325	3	0.50	380	4	0.54	440	5	0.56
2001-3000	115	0	0.31	190	1	0.41	260	2	0.48	385	4	0.58	450	5	0.60	565	7	0.64
3001-4000	115	0	0.31	195	1	0.41	330	3	0.54	450	5	0.63	510	6	0.65	690	9	0.70
4001-5000	195	1	0.41	260	2	0.50	335	3	0.54	455	5	0.63	575	7	0.68	750	10	0.74
5001-7000	195	1	0.42	265	2	0.50	335	3	0.55	515	6	0.69	640	8	0.73	870	12	0.80
7001-10,000	195	1	0.42	265	2	0.50	395	4	0.62	520	6	0.69	760	10	0.79	1050	15	0.86
10,001-20,000	200	1	0.42	265	2	0.51	460	5	0.67	650	8	0.77	885	12	0.86	1230	18	0.94
20,001-50,000	200	1	0.42	335	3	0.58	520	6	0.73	710	9	0.81	1060	15	0.93	1520	23	1.0
50,001-100,000	200	1	0.42	335	3	0.58	585	7	0.76	770	10	0.84	1180	17	0.97	1690	26	1.1

Figure 1 - Reference 1

Reproduced by permission of John Wiley & Sons, Inc. and Copyright (1959) Bell Telephone Laboratories from Sampling Inspection Tables Single and Double Sampling, 2nd Edition by Harold F. Dodge and Harry G. Romig.

8. Choose a sampling plan for the given lot size and estimated PA. Since the sampling plan is designed as a function of the PA, use the estimated PA as the table entry. Remember to obtain revised PA estimates from new data and if possible to select a more economical plan. For one HERL study, there were 7800 source documents. Based on our estimated PA of 0.61% to 0.80%, we would go to the 2% LTPD single sample table, locate the correct PA column and find the sample size for 7800 forms. This would result in the row corresponding to 7001 to 10000 forms being used. A sample size of 760 forms with no more than 10 errors would be used for the study. For the purposes of our plan, the original source form was considered correct and any code difference on the computer file was considered an error.

9. Find the OC curve of the sampling plan. If the operating characteristic (OC) curve is satisfactory, choose the plan. The OC curve for our plan is shown in Figure 2.

10. Select sample units from the lot by a random procedure. A preferred method for accomplishing randomization is the use of random numbers.

11. Follow the prescribed procedure for single sampling. Inspect each unit for the characteristics adopted in step one and in accordance with sampling procedures.

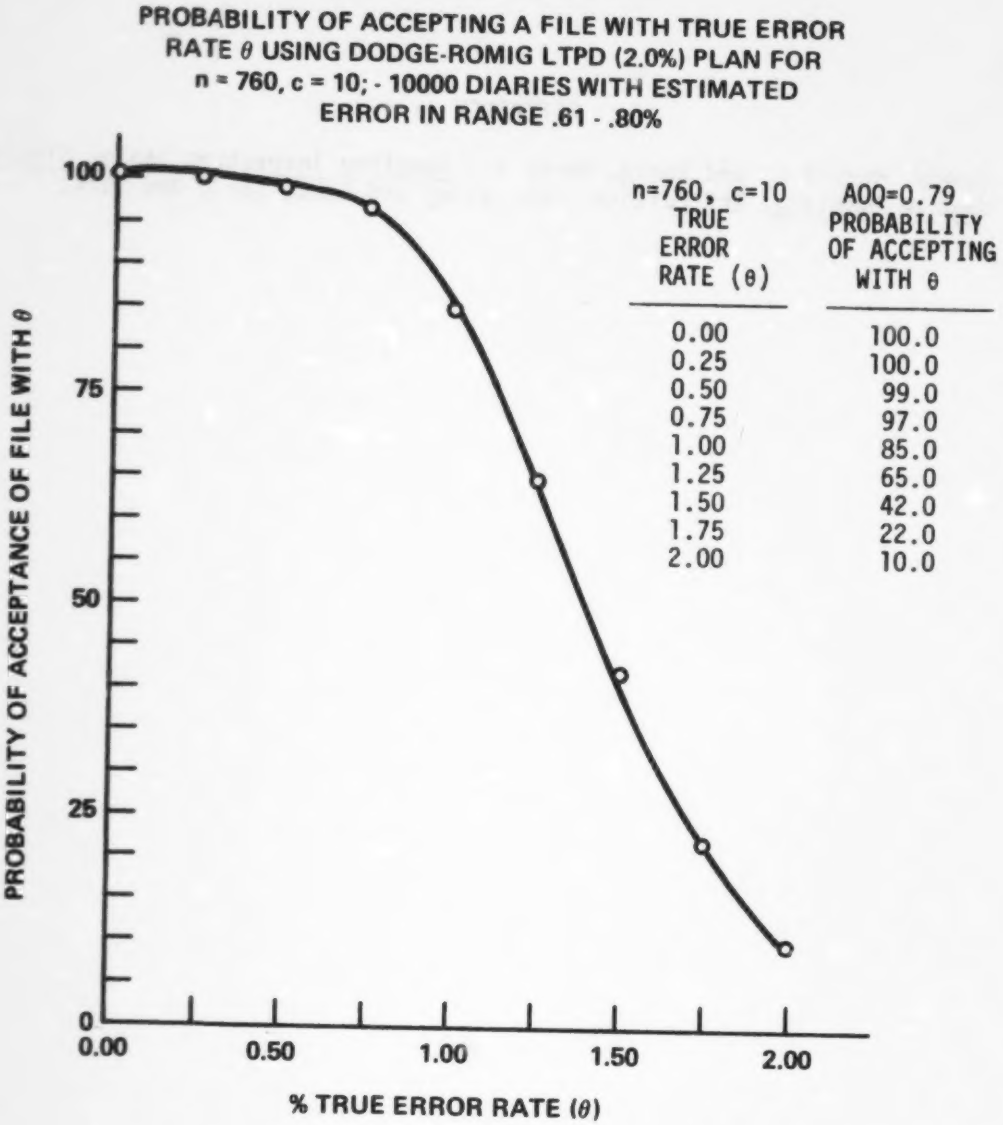
12. Keep a running check of the PA. Change the sampling plan as necessary to match shifts in the PA. Adopt a definite time period for making new estimates such as every month or every quarter. In our experience, the PA did not change significantly over 6 to 7 months.

The Dodge-Romig acceptance sampling plan described is not only being used on the past CHESS studies but is also used on current studies. For each study undertaken by the data processing staff, a data processing protocol is prepared in addition to the normal study protocol. The protocol describes what is to be done including manual and computer steps and the expected timeframe. Edit checks to be performed usually include

1. Check for valid codes
2. range checks
3. field type, numeric, alpha and/or alphanumeric
4. consistency checks such as date of birth versus age.

Edits may be accomplished by an individually designed program or one or more SPSS runs. SPSS frequency distributions are principally used to identify out of range and other unacceptable codes. Audit trails are maintained throughout the processing.

In conclusion, we believe we have a successful operational quality control program for our current needs relative to processing of large computer data files. For similar applications, I would recommend reviewing these procedures as described by Dodge and Romig and investigating more usage of SPSS as a quick evaluation of the contents of the data files.



PROBABILITY OF ACCEPTING A FILE WITH TRUE ERROR RATE θ USING DODGE-ROMIG LTPD (2.0%) PLAN FOR n = 760, c = 10; - 10000 DIARIES WITH ESTIMATED ERROR IN RANGE .61 - .80%

n=760, c=10 TRUE ERROR RATE (θ)	AOQ=0.79 PROBABILITY OF ACCEPTING WITH θ
0.00	100.0
0.25	100.0
0.50	99.0
0.75	97.0
1.00	85.0
1.25	65.0
1.50	42.0
1.75	22.0
2.00	10.0

Figure 2 - Operating Characteristics Curve

Reproduced by permission of John Wiley & Sons, Inc. and Copyright (1959) Bell Telephone Laboratories from Sampling Inspection Tables Single and Double Sampling, 2nd Edition by Harold F. Dodge and Harry G. Romig.

REFERENCE

1. Dodge, Harold F. and Romig, Harry G., Sampling Inspection Tables Single and Double Sampling, 2nd Edition, John Wiley and Sons, Inc., New York, 1959.

USE OF SUCCESSIVE TIME DIFFERENCES AND DIXON RATIO TEST FOR DATA VALIDATION

by

Tyler Hartwell

Research Triangle Institute
Research Triangle Park, North Carolina 27709

USE OF SUCCESSIVE TIME DIFFERENCES AND DIXON RATIO TEST FOR DATA VALIDATION

Tyler Hartwell*

ABSTRACT

This paper describes preliminary work on two statistical data editing procedures designed to flag suspect minute and hourly data from the Regional Air Pollution Study (RAPS) computer data bank which contains data from the Regional Air Monitoring System (RAMS) network of monitoring stations in St. Louis, Missouri. In particular, the data editing procedures are: (i) an intraparameter check where the differences of successive minute averages for a given variable and station are evaluated, and (ii) an intraparameter check where hourly averages for a given hour and variable are compared across the RAMS network or across a selected subset of stations by use of the Dixon ratio. The paper describes how the procedures were developed for their current application and gives results of applying the procedures to actual data on the RAPS data bank. In addition, suggestions for future research on the two procedures are presented. It is concluded that at the present time the two data editing procedures should be useful to EPA in flagging suspect minute and hourly data from the RAPS data bank.

---

* Dr. Hartwell is a senior statistician, Statistical Methodology and Analysis Center, Research Triangle Institute, Research Triangle Park, North Carolina 27709.

I. INTRODUCTION

The RAMS network of 25 monitoring stations in and around St. Louis, Missouri collects data on a large number of pollutant (e.g., O_3, CO, THC, CH_4, NO, NO_x, SO_2, TS, H_2S) and meteorological variables (e.g., wind speed, wind direction, temperature, dew point, delta temperature, barometric pressure). Figure 1 presents a map of the location of the 25 RAMS stations. The figure indicates that the urban stations (nos. 101-108) may be as much as 8 miles apart while the rural stations (e.g., nos. 122-125) may be as much as 35 miles apart. The RAPS Data Bank contains data from the RAMS network of stations.

The purpose of the two statistical data editing rules (i.e., minute successive differences and the Dixon Ratio) examined in this paper is only to flag suspect RAPS data, not to delete it from the data bank. That is, because of the vast amount of data collected by the RAMS network, data editing rules are needed to limit the amount of suspect data that meteorologists and atmospheric chemists need to examine in detail. Thus, the purpose of this paper is to examine two data editing rules that indicate data that should be examined in more detail by EPA personnel who have an intimate knowledge of the data that the RAMS network collects.

In addition, it is important to note here that the work presented here is only preliminary. Because of the complexity of trying to obtain data editing rules that apply to a large network of monitoring stations, additional work needs to be done on refining the two rules. However, at this point in time, it is felt that the two data editing rules presented should prove to be useful in flagging suspect data from the RAMS network.

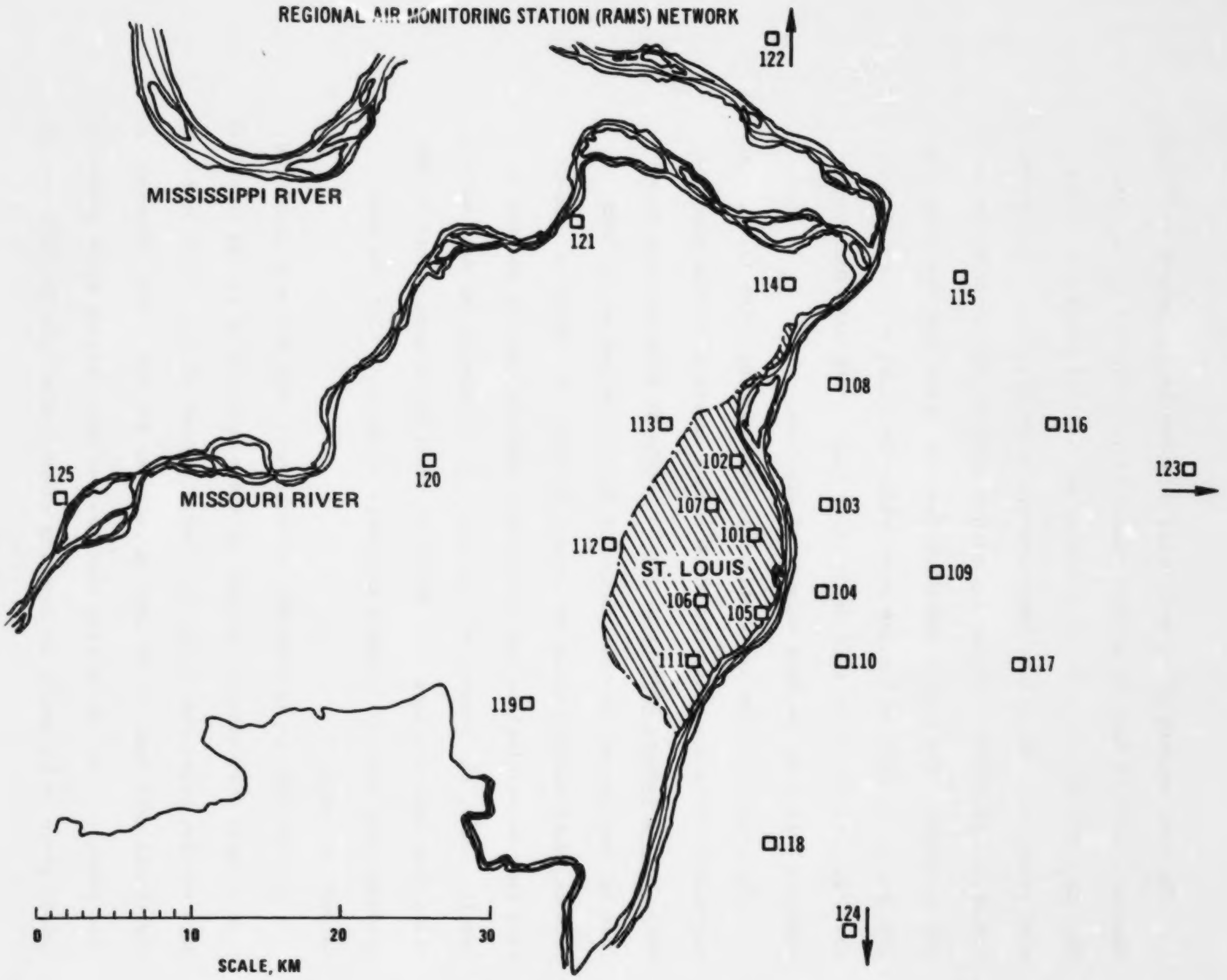
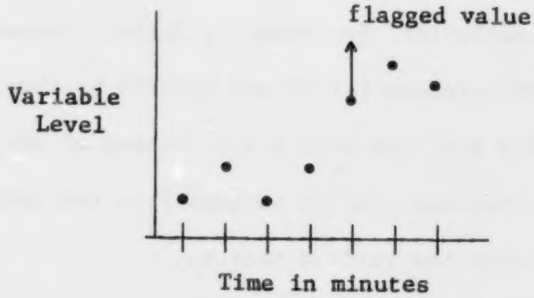

REGIONAL AIR MONITORING STATION (RAMS) NETWORK
MISSISSIPPI RIVER
MISSOURI RIVER
ST. LOUIS
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
0
10
20
30
SCALE, KM

Figure 1. RAMS Network of 25 Stations

II. MINUTE SUCCESSIVE DIFFERENCES

The RAMS data received at the Research Triangle Park, North Carolina contains minute data on several air pollution and meteorological variables. Several computerized range validation checks are performed on this data by the prime contractor, prior to forwarding it to the RAPS Data Bank. The RAPS Data Bank was interested in determining if a statistical procedure could be used in further validation of the data, to flag minute data values which appeared to be outliers. In particular, there was a need to develop and evaluate a procedure (i) which could be applied to each station's data for one variable at a time and (ii) was easy to compute and only required one pass through the data. Accordingly, this study was limited to a simple statistical procedure that required little computation. After discussions between EPA and RTI staff members, it was decided to examine a statistical data editing rule based on minute successive differences.

In general, the editing rule examined is designed to flag minute values which are relatively much higher or lower than the preceding minute value; i.e.,



Thus, the editing rule is designed to detect large spikes in the minute values of a variable at a station.

In particular, the data editing rule is the following: at a particular station compute successive differences between minute values of a particular variable and if a successive difference is "too large" then flag this value. This rule is extremely simple to apply and requires only one pass through the data base.

In order to determine when a successive difference was "too large", the (i) distributions, (ii) sample means, and (iii) sample standard deviations (s.d.) of minute successive differences for several stations, times of the day, and air pollution and meteorological variables from the RAMS network were examined. For example, Figures 2, 3, and 4 present three of these distributions for the variables windspeed, ozone, and NO_2. In all, over 200 of these minute successive difference plots were examined.

After examining these distributions and the corresponding sample means and standard deviations in detail, it appeared reasonable to assume that in general the minute successive differences were approximately normally distributed with a mean of zero. However, it was also clear that the standard deviation of minute successive differences was not constant over stations, times of the day, seasons of the year, and pollutant or meteorological variables. For example, Table 1 presents s.d.s of minute successive differences for CO and methane by the factors time of the day (0-4 a.m., 4-8 a.m. and 8-12 a.m.), season of the year, and two rural and two urban stations. It is obvious from the table that the s.d.s vary a great deal over the various factors.

Accordingly, it was decided to assume that the distribution of minute successive differences for variables in the RAMS network was normally distributed with a mean of zero and a standard deviation that

FIGURE 2

DISTRIBUTION OF WINDSPEED MINUTE SUCCESSIVE DIFFERENCES
STATION 101; DAY 190, 1976; TIME 0 TO 4 A.M.

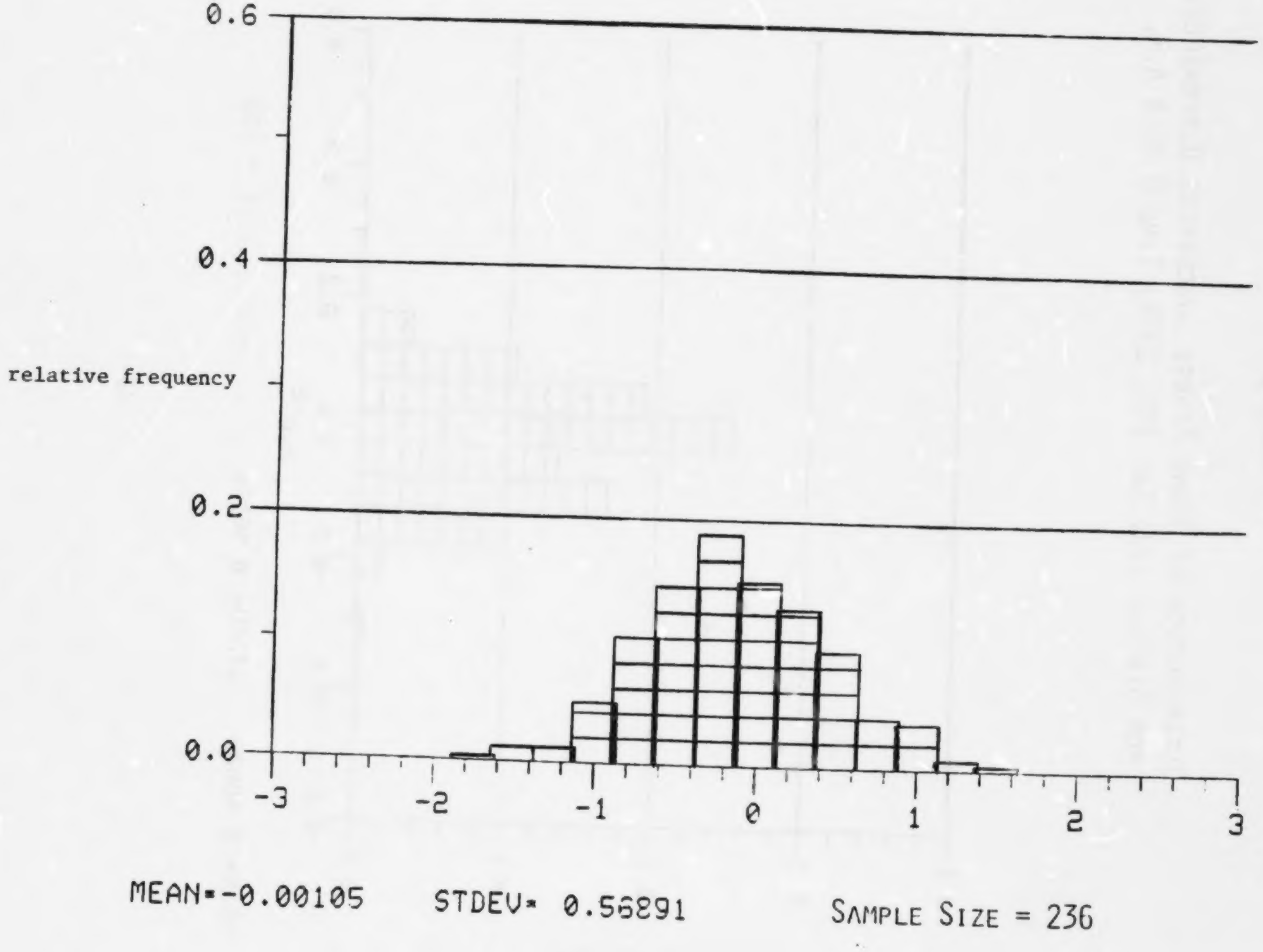
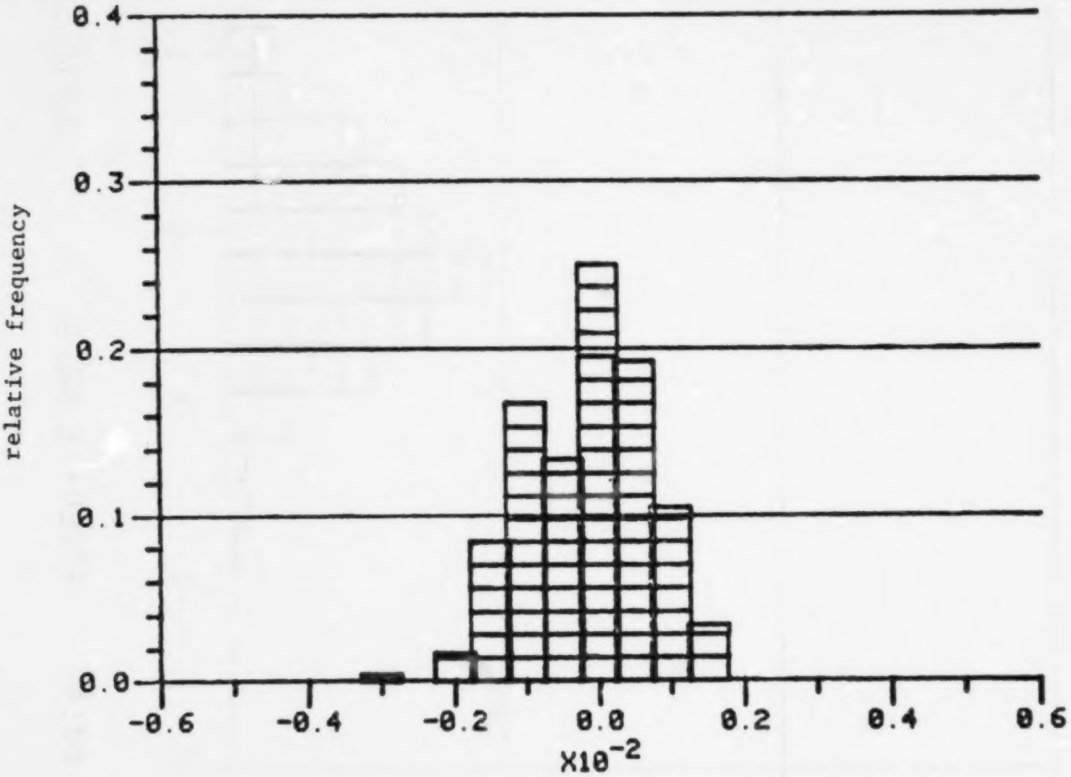


FIGURE 3

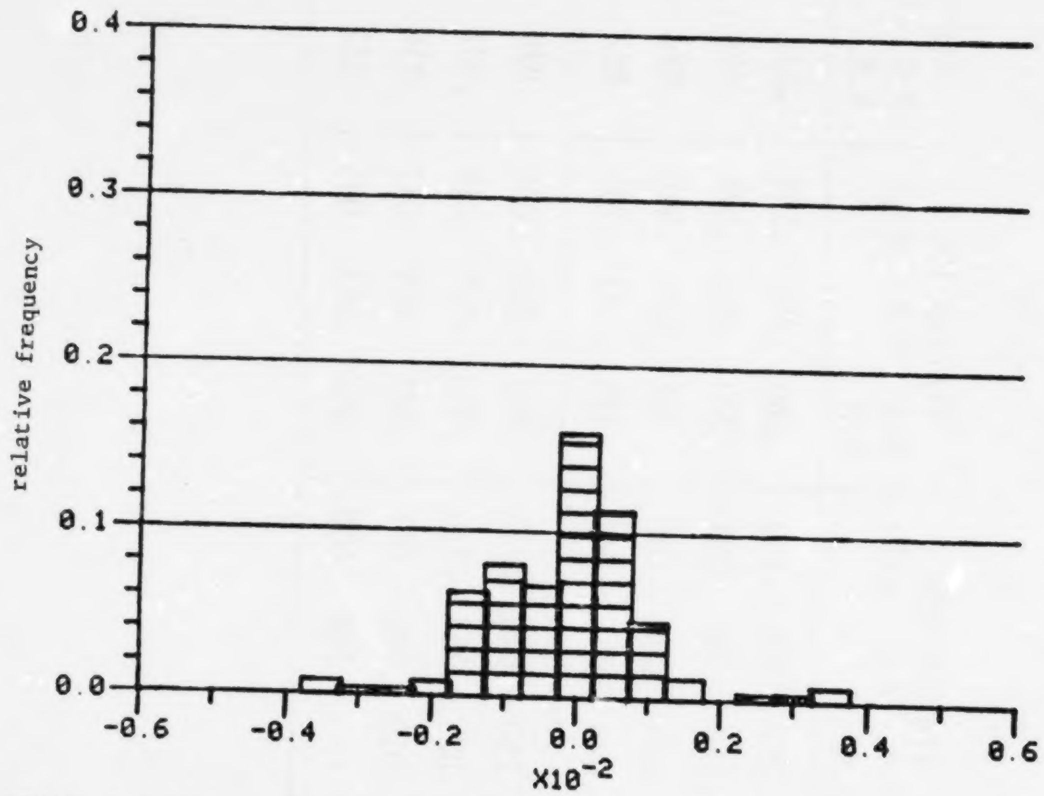
DISTRIBUTION OF OZONE MINUTE SUCCESSIVE DIFFERENCES FOR STATION 122; DAY 180, 1976; TIME 0 TO 4 A.M.



MEAN= 0.00009 STDEV= 0.00084 SAMPLE SIZE = 236

FIGURE 4

DISTRIBUTION OF NO_2 MINUTE SUCCESSIVE DIFFERENCES FOR STATION 122; DAY 180, 1976; TIME 4 TO 8 A.M.



MEAN= 0.00007 STDEV= 0.00109

SAMPLE SIZE = 139

TABLE 1

STANDARD DEVIATIONS OF MINUTE SUCCESSIVE DIFFERENCES IN 1976 FOR CO AND METHANE BY TIME OF DAY, SEASON OF YEAR, AND TYPE OF STATION

VARIABLE	DAY/SEASON	STATION 101			STATION 104			STATION 117			STATION 122		
		0-4 A.M.	4-8	8-12	0-4 A.M.	4-8	8-12	0-4 A.M.	4-8	8-12	0-4 A.M.	4-8	8-12
CO	40/WINTER	.252	.165	.149	.260	.151	.109	.091	.037	.119	----	----	----
	129/SPRING	.054	.394	.081	.270	.283	.522	.092	.031	.464	.014	.028	.030
	220/SUMMER	.015	.286	.046	.122	.231	.416	.029	.026	.036	.014	.017	.010
	270/FALL	.045	.091	.067	.189	.091	.295	.027	.023	.021	.013	.067	.022
METHANE	40/WINTER	.160	.040	.033	.061	.033	.006	.024	.017	.019	----	----	----
	129/SPRING	.025	.018	.010	.014	.028	.010	.007	.009	.008	.023	.017	.011
	220/SUMMER	.010	.009	.007	.020	.014	.017	.016	.013	.028	.005	.006	.005
	270/FALL	.046	.014	.008	.052	.043	.035	.012	.016	.014	.006	.008	.007

- = INSUFFICIENT DATA.

may vary over time of the day, season, and type of station (urban or rural). This implied that a minute successive difference would be flagged when it was greater than a function of the appropriate standard deviation (e.g., a standard procedure for detecting outliers for the normal distribution with mean zero is to flag observations which are greater (or less) than 4 s.d.s). The probability of an observation being greater (or less) than 4 s.d.s for the normal distribution is less than .0001. Thus, the problem reduced to determining for each variable of interest the appropriate standard deviation which might depend on time of the day, season of the year, and type of station.

To examine this problem, standard deviations of minute successive differences for each variable of interest for approximately 4 days per season in 1976, 4 stations (2 urban and 2 rural), and 3 times of the day (0-4 a.m., 4-8, 8-12) were computed. Thus for each variable between 100 and 192 (4 days x 4 seasons x 4 stations x 3 times = 192) standard deviations were computed. A standard deviation was only computed if at least 60 minute successive differences were available during the 4-hour time period being considered. The results of some of these computations are summarized in Table 2 for 10 of the variables measured in the RAMS network. These 10 variables were chosen not only because there was interest in editing their minute values but also because sufficient data was available on them for the days selected for computing standard deviations. The s.d.s presented in Table 2 are average s.d.s (i.e., averaged over several stations and days).

Using the s.d.s in Table 2 a statistical technique referred to as the analysis of variance (ANOVA) was used to test if these average standard deviations were significantly different (in a statistical

Table 2. Sample Means of Standard Deviations of Minute Successive Differences for RAMS Data; by Variable[1/]

Variable	Station		Time of Day (A.M.)			Season of Year				Overall[2/] Average
	Urban	Rural	0-4	4-8	8-12	Winter	Spring	Summer	Fall	
Windspeed (meters/sec.)	.59	.47	.41	.44	.75	.58	.64	.52	.41	.53
Temperature (^oC)	.10	.08	.06	.06	.17	.10	.13	.09	.06	.09
Ozone (ppm)	.0020	.0012	.0012	.0012	.0021	.0007	.0021	.0024	.0012	.0016
CO (ppm)	.1 1	.117	.077	.119	.222	.096	.105	.150	.206	.139
Methane (ppm)	.031	.019	.029	.028	.019	.024	.024	.017	.035	.025
THC (ppm)	.094	.055	.056	.084	.084	.065	.057	.082	.094	.075
NO (ppm)	.0069	.0011	.0032	.0048	.0039	.0036	.0042	.0031	.0050	.004
NO_X (ppm)	.0088	.0020	.0045	.0062	.0055	.0058	.0059	.0040	.0059	.0054
Total Sulfur (ppm)	.0027	.0009	.0015	.0015	.0024	.0028	.0010	.0002	.0032	.0018
SO_2 (ppm)	.0015	.0006	.0010	.0010	.0011	.0022	.0011	.0003	.0005	.0011

[1/] Each sample mean based on S.D.s over several stations and days (stations = 101, 104, 117, and 122; days = 39, 40, 61, 100, 129, 130, 170, 180, 219, 220, 265, 270, 299, 300, 305, 319.

[2/] Overall Average = mean over time of day, season, and type of station.

sense). In particular, a separate ANOVA was carried out for each of the 10 variables. In each of the 10 ANOVAs, statistical tests were used to determine if average standard deviations by time of day, type of station, and season of the year were significantly different. The results of running the 10 ANOVAs indicated that in the majority of cases the average s.d.s in Table 2 were significantly different. For example, in column 3 of the table (i.e., ozone by station) the average s.d.s of minute successive differences for ozone for two urban stations was .0020 and two rural stations was .0012. The test of significance of these two averages was significantly different at the .01 level. Note that Table 2 also presents the average s.d. for each variable over time of day, season of the year, and type of station (e.g., .0016 for ozone).

The average standard deviations in Table 2 clearly indicate that from a statistical point of view the s.d.s of minute successive differences are significantly different for several of the variables for one or more of the factors examined (i.e., station type, time of day, and season of the year). Thus, to be strictly correct (in a statistical sense) in applying the data editing rule based on minute successive differences, it would be necessary to base the rule on varying s.d.s by season of the year, etc. (i.e., the rule would be ±4 s.d.s where the s.d.s are given in Table 2).

Due to the fact that the above data editing rule might prove to be somewhat confusing, a more conservative and easier to program rule has been initially examined (of course, only actual application of a data editing rule can determine its practical usefulness). This rule is based upon using, for all minute successive differences for each variable in Table 2, ±4 times the largest average s.d. across station type, time of day, and season of the year. Thus, for ozone the rule would be based on ±4 times the s.d. = .0024 (i.e., the average s.d. for summer). This rule is extremely easy to apply requiring only one value to be exceeded by each minute successive difference of a variable regardless of station, time of day, or season. Of course, it is conservative in the sense of having a limit which is somewhat high in many cases (e.g., for ozone in winter a more exact rule would be based on a s.d. = .0007).

Accordingly, using the largest average s.d. for each variable in Table 2 and basing the data editing rule on ±4 s.d. limits, Table 3 gives possible limits for flagging minute successive differences by variable for the RAMS network of stations. In deriving the limits in Table 3, it was noted for the RAMS network that CO, methane, and THC were only measured every 5 minutes and total sulfur (with the exception of Station 117) and SO_2 were only measured every 3 minutes. Therefore, the s.d.s given to RTI based on minute successive differences of these variables were underestimates for detecting spikes at five (or three) minute intervals (i.e., the average s.d.s in Table 2 are too small for these five variables). In an attempt to compensate for this underestimate, Table 3 includes an adjustment factor for these five variables. The adjustment factor multiplies the ±4 s.d. limits for CO, methane, and THC by $\sqrt{5}$ and the ±4 s.d. limits for total sulfur and SO_2 by $\sqrt{3}$. These factors were derived by assuming that on the RAMS data file the minute successive differences for CO, methane, and THC are zero except for every 5th minute (etc. for total sulfur and SO_2).

Using the limits given in Table 3, RTI then examined the percentage of minute successive differences that would be flagged for Stations 101 and 122 for 8 days in 1976 for 10 RAMS variables. The results of these

TABLE 3

POSSIBLE MINUTE SUCCESSIVE DIFFERENCE LIMITS ON 10 RAMS VARIABLES[1/]

RAMS VARIABLE	LIMIT
WINDSPEED (METERS/SEC.)	±3.0
TEMPERATURE (°C)	± .660
OZONE (PPM)	± .0096
CO (PPM)	±1.97
METHANE (PPM)	± .316
THC (PPM)	± .844
NO (PPM)	± .028
NO_x (PPM)	± .035
TOTAL SULFUR (PPM)	± .022
SO_2 (PPM)	± .015

[1/] BASED ON ±4 STANDARD DEVIATION LIMITS. IN ADDITION, FOR CO, METHANE, AND THC, THE ±4 S.D. LIMITS HAVE BEEN MULTIPLIED BY $\sqrt{5}$ TO ADJUST FOR THE FACT THAT THESE VARIABLES ARE ONLY MEASURED EVERY 5 MINUTES. SIMILARLY, FOR TOTAL SULFUR AND SO_2 THE ±4 S.D. LIMITS HAVE BEEN MULTIPLIED BY $\sqrt{3}$ SINCE THESE VARIABLES ARE ONLY MEASURED EVERY 3 MINUTES.

computations are given in Table 4. The table shows that except for ozone the percent flagged per variable was less than .6 percent. For ozone it appears that entirely too many minute differences were flagged. For example, Figure 5 presents a plot over two days of minute values of ozone for Station 105. Examination of the figure indicates that a relatively large percentage of the minute values would be flagged using the limits given in Table 3; although, EPA personnel have indicated that the data in Figure 5 is not atypical. In addition, discussions with EPA personnel have indicated that in the RAMS network, five of the stations (101, 104, 105, 107, and 115) are heavily affected by traffic. Thus, it may be necessary for these stations to have higher limits for flagging minute values for ozone than those given in Table 3. Also, it has been suggested that for these traffic affected stations it may be necessary to examine both ozone and NO_x minute values simultaneously before flagging ozone values (e.g., if a minute ozone value jumps significantly from one minute to the next but the NO_x reading does not jump, then and only then should the ozone value be flagged.)

In addition to the results given in Table 4, the percentage of minute successive differences that would be flagged for Station 105 for 8 days in 1976 using the limits given in Table 3 were examined. The results were the following:

Percentage Flagged for Station 105

Variable	Percent Flagged
WS	.04
Temp.	.12
O_3	7.9
CO	.56
CH_4	.41
THC	.90
NO	1.5
NO_x	1.4
TS	1.5
SO_2	1.8

The above table clearly indicates that additional work needs to be done for ozone since entirely too many values are being flagged. In addition, for the other pollutant variables many more values are being flagged than in Table 4 (Stations 101 and 122 combined). Thus, it would seem that the variables at Station 105 are probably being affected by automobile traffic.

Accordingly, at the present time it appears that the minute successive differences limits in Table 3 except for ozone are probably reasonable for a majority of the RAMS monitoring stations. However, for Stations 101, 104, 105, 107, and 115 which are heavily affected by traffic, additional work needs to be done on the limits for the pollutant variables (the limits for windspeed and temperature appear reasonable for all stations). Of course, since only Stations 101, 104, 105, 116, 117, and 122 were examined in this analysis to date, it may be that additional work is needed for other specific stations. For Stations 101, 104, 105, 107, and 115 wider minute successive differences limits should be examined for the pollutant variables particularly for ozone. As mentioned previously, it may be for ozone that simple minute successive difference limits are impractical (i.e., for ozone at the traffic effected stations it may be necessary to have a minute data editing rule which is tied to another pollutant such as NO_x).

Another refinement of the limits given in Table 3 that might be examined is to have them vary by time (see Table 2). For example, for ozone the s.d. of minute successive differences is much higher for 8-12 a.m. than from 0-8 a.m.

Finally, before proceeding, Figures 6 and 7 present two plots of minute values for CO from the RAPS Data Bank for Stations 101 and 105, respectively. These plots indicate minute CO values which were flagged by the CO limits given in Table 3. The plots indicate that the data editing rule based on minute successive differences may be quite useful in detecting minute outliers for CO.

TABLE 4

PERCENTAGE OF MINUTE SUCCESSIVE DIFFERENCES FLAGGED FOR STATIONS 101 AND 122 FOR 8 DAYS IN 1976; BY VARIABLE[1/]

VARIABLE	PERCENT FLAGGED
WINDSPEED	.53
TEMPERATURE	.17
OZONE	3.0
CO	.06
CH_4	.11
THC	.24
NO	.12
NO_x	.12
TS	.31
SO_2	.34

[1/] TOTAL NUMBER OF MINUTE SUCCESSIVE DIFFERENCES COMPUTED = 23,040 (8 DAYS x 2 STATIONS x 1,440 MINUTES/DAY). THE DAYS WERE 1, 2, 96, 97, 231, 232, 286 AND 287.

FIGURE 5

O_3 VERSUS TIME BY MINUTE FOR TWO DAYS IN 1976

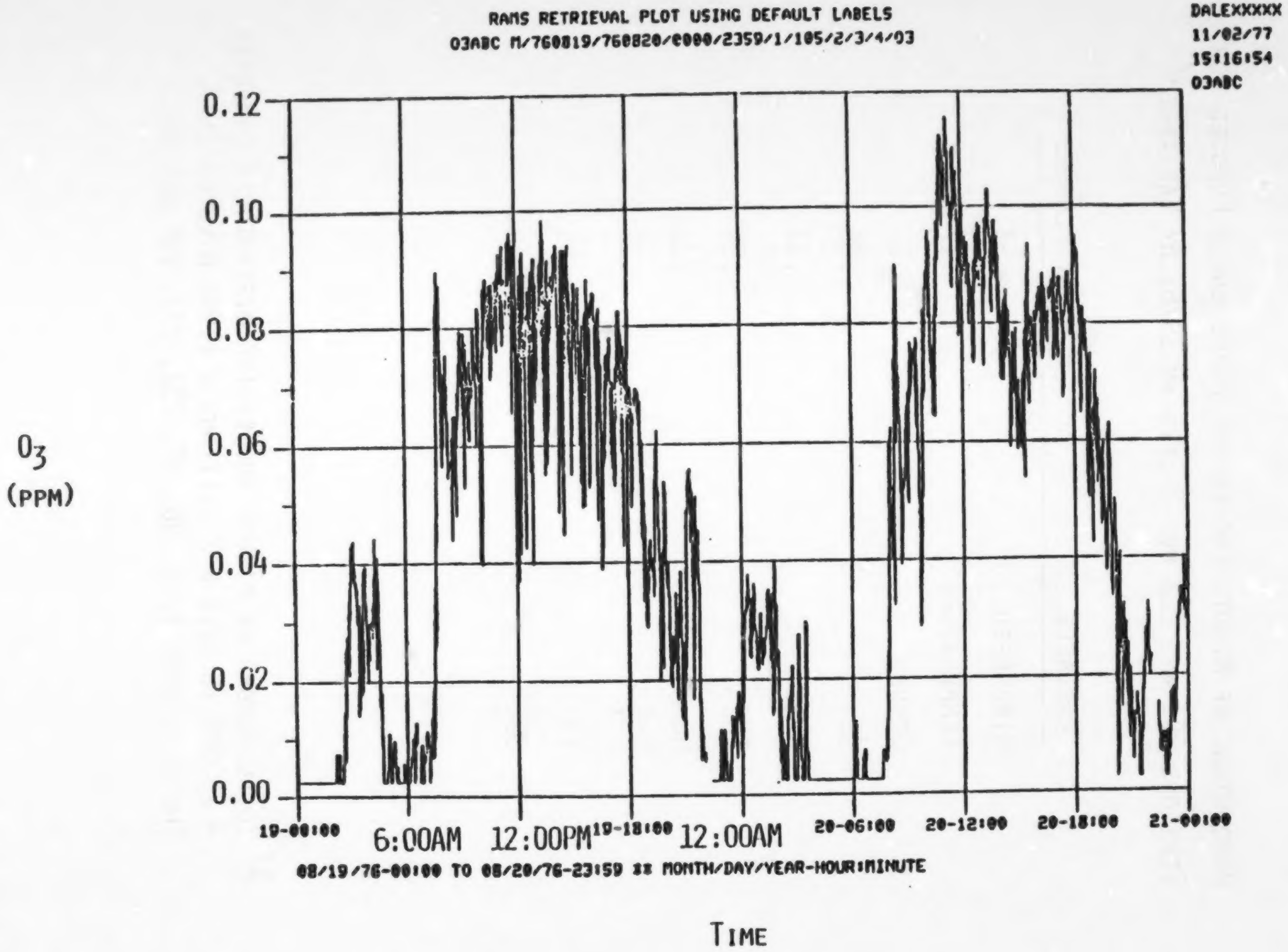


FIGURE 6

CO Versus Time by Minute for Day 1 in 1976

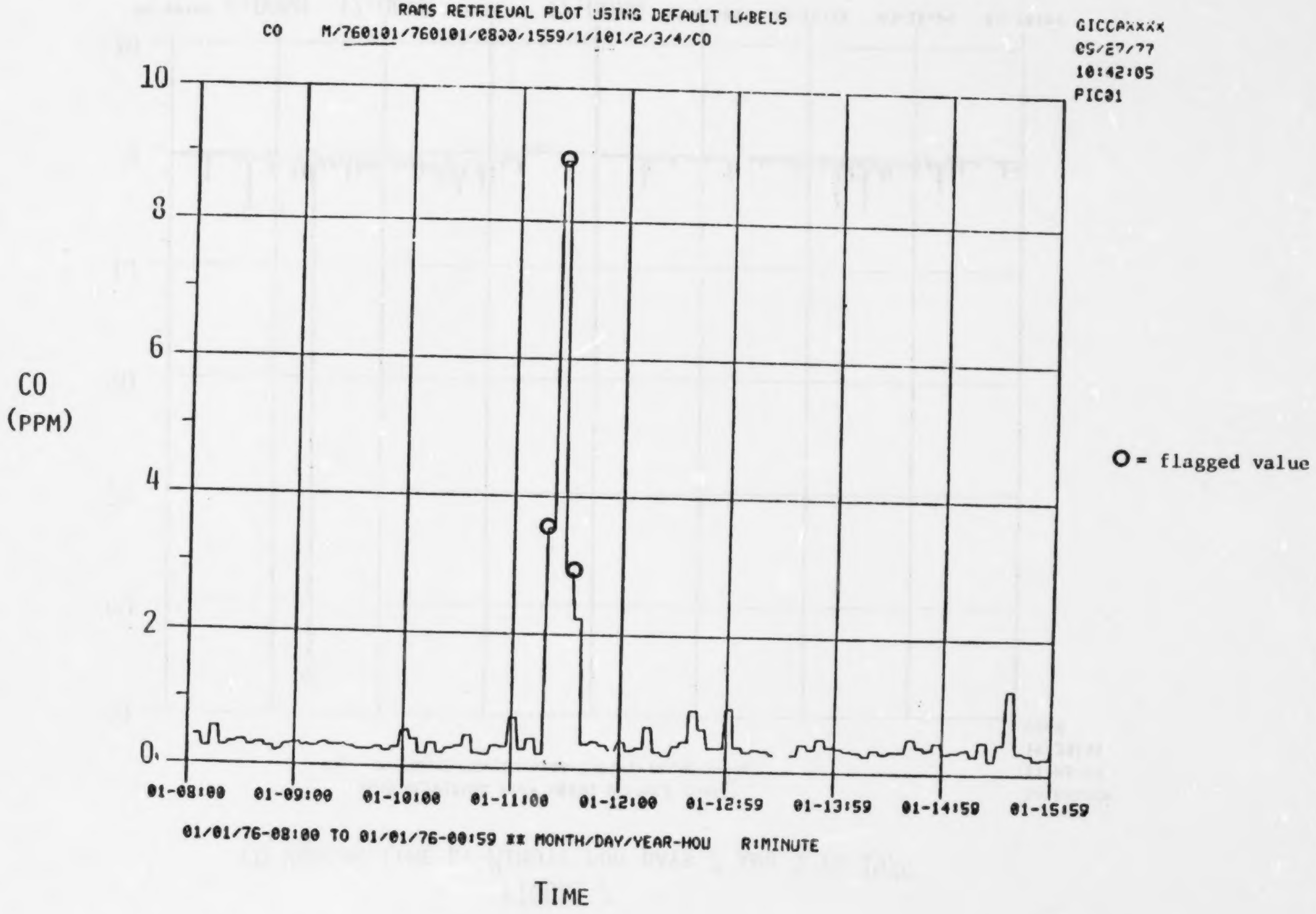
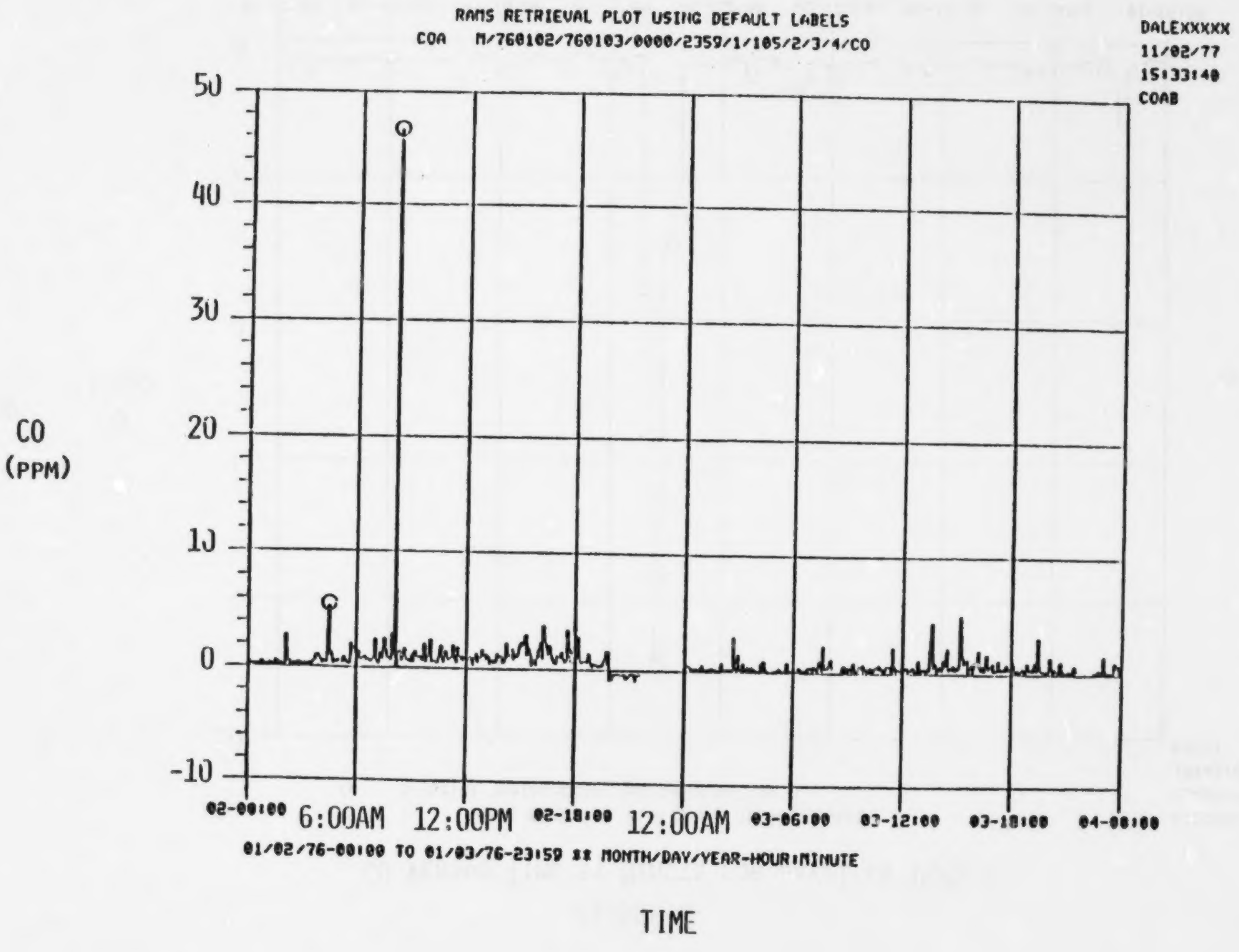


FIGURE 7

CO VERSUS TIME BY MINUTE FOR DAYS 2 AND 3 IN 1976



III. DIXON RATIO

This study also examined the possibility of flagging hourly data across the RAMS network by use of the Dixon Ratio. That is, for a particular variable and hour of the day this ratio will flag stations whose hourly averages are "too high" or "too low" as compared with the other stations in the network for that hour and variable. The Dixon criterion was examined as a potential validation procedure for hourly averages because it was a simple procedure which was easy to compute and only required one pass through the data.

In brief the Dixon criterion is the following:

For a particular hour and variable (e.g., O_3) rank the N hourly values X_i over stations such that $X_1 \leq X_2 \leq \ldots \leq X_N$. Then compute the criterion (if N is between 11 and 13)

$$R_H = \frac{X_N - X_{N-2}}{X_N - X_2} \quad \text{(to check largest hourly value)}$$

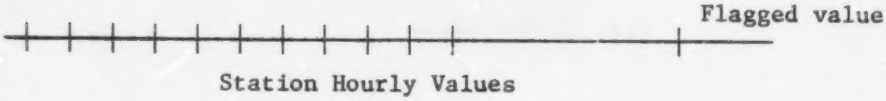
$$R_L = \frac{X_3 - X_1}{X_{N-1} - X_1} \quad \text{(to check smallest hourly value)} \qquad (1)$$

If R_H (or R_L) is too large reject largest (smallest) hourly value (e.g., if number of stations = 12 and $R_H > .642$, reject X_N at the .01 level of significance under normal distribution assumptions). Note, see [1] for the Dixon criterion when N is less than 11 or greater than 13.

In general, the Dixon criteria is designed to reject the following type of station hourly values:

---

[1] Dixon, W. J., "Processing Data for Outliers," Biometrics, Vol. 9 (1953), p. 74.



Thus, the criteria is designed to flag hourly values which contribute a relatively large percentage of the range of the hourly values across the network. Note, that the criterion as given in (1) only flags one high and one low hourly value and not multiple hourly values. For the present study RTI did not examine the flagging of multiple hourly values for a particular hour of the day.

The Dixon criterion given in Equation (1) was first applied across the entire RAMS network of 25 stations. The results of these calculations indicated that entirely too many hourly values were being flagged for several of the RAMS variables (e.g., ozone, CO, and NO_2). Accordingly, the reason so many hourly values were being flagged was determined by examining the sample means of several RAMS variables for urban, residential, and rural stations in the RAMS network. Table 5 gives a summary, for six RAMS variables, of some of these computations. The table gives the sample means and standard deviations for the three types of stations by season of the year. In general the table shows that the sample means of hourly values for a particular variable are not the same for urban, residential, and rural stations in the RAMS network. (This is particularly true of the pollutant variables; whereas, for the meteorological variables the means are much more similar across the network.) Statistical tests of these station type means were found to be significantly different in several cases. Thus, it became clear that one of the underlying assumptions of applying the Dixon criterion was being violated; namely, that the hourly station values come from a

TABLE 5

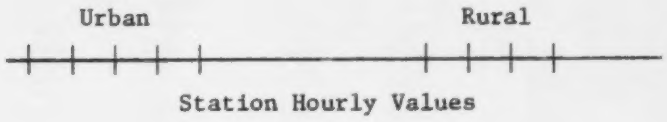
HOURLY SAMPLE MEANS AND STANDARD DEVIATIONS FOR URBAN, RESIDENTIAL, AND RURAL STATIONS[1] IN THE RAMS NETWORK BY SEASON OF THE YEAR AND VARIABLE

		STATION TYPE					
		URBAN		RESIDENTIAL		RURAL	
VARIABLE	SEASON	MEAN[2]	STD. DEV.	MEAN	STD. DEV.	MEAN	STD. DEV.
O_3 (PPM)	SPRING	.030	.013	.034	.013	.041	.015
	SUMMER	.052	.023	.055	.018	.063	.021
	FALL	.017	.021	.018	.011	.025	.013
	WINTER	.008	.006	.009	.006	.016	.008
CO (PPM)	SPRING	.699	.756	.634	.670	.333	.474
	SUMMER	.999	.917	1.072	.837	.329	.638
	FALL	.837	.987	.793	1.092	.249	.307
	WINTER	.451	.526	.418	.327	.215	.163
NO_2 (PPM)	SPRING	.030	.017	.025	.016	.011	.012
	SUMMER	.046	.087	.047	.094	.011	.010
	FALL	.026	.015	.023	.012	.011	.010
	WINTER	.024	.012	.017	.010	.010	.008
CH_4 (PPM)	SPRING	1.781	.257	1.604	.202	1.480	.294
	SUMMER	2.068	.486	1.905	.279	1.778	.297
	FALL	1.834	.391	1.787	.244	1.621	.195
	WINTER	1.768	.275	1.436	.235	1.604	.210
WINDSPEED (METERS/SEC.)	SPRING	4.122	1.775	3.928	1.525	4.020	1.839
	SUMMER	2.799	1.082	2.710	1.183	2.249	.898
	FALL	4.518	1.385	4.122	1.252	3.967	1.483
	WINTER	5.255	1.656	4.871	1.602	5.277	2.097
TEMPERATURE (°C)	SPRING	13.993	4.337	13.864	4.314	13.367	4.408
	SUMMER	26.474	1.879	25.767	3.153	25.497	2.026
	FALL	11.924	6.271	11.277	6.212	10.786	6.335
	WINTER	-5.180	6.648	-4.612	7.228	-5.535	6.722

[1] URBAN = STATIONS 101 TO 108; RESIDENTIAL = STATIONS 111 TO 113, 119, AND 120; RURAL = STATIONS 109, 110, 114 TO 118, 121 TO 125.

[2] MEANS AND STANDARD DEVIATIONS ARE BASED UPON 4 HOURS PER DAY FOR 5 DAYS OVER THE VARIOUS STATIONS (URBAN, RESIDENTIAL, OR RURAL).

normal distribution with the same mean and variance. Instead, for example, the hourly station values for ozone have different means for urban and rural stations (the means are higher for rural stations). The consequence of groups of stations having different means is illustrated below:



The above figure shows, using the Dixon criterion, that some rural stations may be flagged simply because their means are always higher than the means for urban stations. Accordingly, after examining sample means such as those presented in Table 5, RTI decided to apply the Dixon criterion separately to rural stations (Stations 109, 110, 114 to 118, 121 to 125) and urban-residential stations (101-108, 111 to 113, 119, and 129).

After applying the Dixon criterion to the two types of stations separately, RTI then examined the results and again found that too many hourly values were being flagged. Accordingly, after discussions with meteorologists and air chemists it was decided that the Dixon criterion for certain pollutant variables should only be applied across the rural or urban-residential stations if the following criteria were met:

(i) the high station value > twice the low station value, and

(ii) the high station value > some constant (e.g., constant = .03 ppm for ozone).

The first criteria simply means that a factor of 2 for an hourly average across the network is not uncommon. Criteria (ii) limits the application of the Dixon Ratio to situations where most of the measurements are well above minimum detectable. In addition, it was decided that the Dixon criterion could not be used for hourly NO, NO_x, TS, and SO_2 values. Furthermore, it was felt that the use of the Dixon criterion for CO was questionable due to the heavy influence of traffic on this variable.

With the above restrictions, the Dixon criterion for detecting high hourly values only (i.e., R_H in Equation (1)) was then applied to 7 variables in the RAPS Data Bank for both urban-residential and rural stations. In applying the rule an hourly value was flagged if R_H was greater than .7 (except for dew point where $R_H > .6$ was flagged). The results of these computations are presented in Tables 6 and 7. In addition, Tables 8 and 9 present examples of flagged hourly values for several different RAMS variables. Tables 6 and 7 indicate that the percent flagged is usually 5% or less. Thus, the Dixon rule as applied does not seem to be impractical. In addition, the examples given in Tables 8 and 9 clearly indicate hourly values for several variables on the data bank that should be examined in more detail by knowledgeable meteorologists and air chemists.

Accordingly, as with minute successive differences, RTI feels that the Dixon criterion will be useful in flagging hourly data across the RAMS network. However, further refinement of the rule may be required. For example, two points that need further examination are:

(i) can the rule be applied to flagging low hourly values (R_L in Equation (1)), and

(ii) can the rule be applied in practice to the stations which are heavily influenced by traffic (101, 104, 105, 107, and 115), particularly for the variable, CO? An alternative here would be to only use the Dixon rule for the 20 stations not heavily influenced by traffic.

TABLE 6

RESULTS OF APPLYING DIXON RATIO TO 12 URBAN STATIONS IN RAMS NETWORK[1/][2/][3/]

VARIABLE	PERCENT OF TIME RATIO > .7	NUMBER FLAGGED
OZONE	1.0	5
CO	2.9	14
CH_4	2.1	10
THC	5.2	25
TEMPERATURE	2.7	13
DEW POINT[4/]	4.4	21
WINDSPEED	4.0	19

[1/] RATIO APPLIED TO ALL 24 HOURS ON 20 DIFFERENT DAYS FOR THE VARIOUS POLLUTANT AND METEOROLOGICAL VARIABLES (= 480 RATIOS FOR EACH VARIABLE).

[2/] RATIO ONLY USED TO FLAG HIGH VALUES.

[3/] FOR OZONE RATIO APPLIED ONLY IF HIGH STATION VALUE > .03 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR CO RATIO APPLIED ONLY IF HIGH STATION VALUE > 3.0 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR CH_4 RATIO APPLIED ONLY IF HIGH STATION VALUE > 2.0 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR THC RATIO APPLIED ONLY IF HIGH STATION VALUE > 2.0 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR WINDSPEED RATIO APPLIED ONLY IF HIGH STATION VALUE > 3.0 METERS SEC.

[4/] FOR DEW POINT PERCENT OF TIME RATIO > .6.

TABLE 7

RESULTS OF APPLYING DIXON RATIO TO 13 RURAL STATIONS IN RAMS NETWORK[1/][2/][3/]

VARIABLE	PERCENT OF TIME RATIO > .7	NUMBER FLAGGED
OZONE	3.8	18
CO	6.0	29
CH_4	4.2	20
THC	5.0	24
TEMPERATURE	8.5[5/]	41
DEW POINT[4/]	6.5	31
WINDSPEED	4.4	21

[1/] RATIO APPLIED TO ALL 24 HOURS ON 20 DIFFERENT DAYS FOR THE VARIOUS POLLUTANT AND METEOROLOGICAL VARIABLES (= 480 RATIOS FOR EACH VARIABLE).

[2/] RATIO ONLY USED TO FLAG HIGH VALUES.

[3/] FOR OZONE RATIO APPLIED ONLY IF HIGH STATION VALUE > .03 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR CO RATIO APPLIED ONLY IF HIGH STATION VALUE > 3.0 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR CH_4 RATIO APPLIED ONLY IF HIGH STATION VALUE > 2.0 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR THC RATIO APPLIED ONLY IF HIGH STATION VALUE > 2.0 PPM AND HIGH STATION VALUE > 2 LOW STATION VALUE.
FOR WINDSPEED RATIO APPLIED ONLY IF HIGH STATION VALUE > 3.0 METERS SEC.

[4/] FOR DEW POINT PERCENT OF TIME RATIO > .6.

[5/] FOR TEMPERATURE THE DIXON RATIO FLAGGED EVERY HOURLY VALUE FOR TWO CONSECUTIVE DAYS IN WINTER WHERE ONE STATION IN THE RURAL NETWORK READ APPROXIMATELY 8°C AND ALL OTHER STATIONS READ LESS THAN -2°C, SEE TABLE 9 (THIS COULD PERHAPS BE DUE TO A SIGN MISTAKE AT THE HIGH STATION).

TABLE 8

EXAMPLES OF FLAGGED VALUES FOR SEVERAL VARIABLES USING DIXON RATIO ON 12 URBAN STATIONS IN THE RAPS DATA BANK[1/]

VARIABLE	DIXON RATIO	STATION VALUES					
		HIGHEST	2 HI	3 HI	3 LO	2 LO	LOWEST
WINDSPEED (METERS/SEC.)	.895	15.7	4.5	3.6	2.4	2.1	2.0
	.821	14.4	4.3	4.2	2.6	1.9	1.7
	.776	6.4	2.7	2.6	1.5	1.5	1.3
TEMPERATURE (°C)	.766	6.5	3.2	1.0	-.30	-.65	-1.6
	.768	32.1	29.2	28.4	27.3	27.2	27.0
	.769	18.7	15.9	15.2	14.2	14.2	13.9
DEW POINT	.724	-2.8	-6.2	-9.6	-11.8	-12.1	-13.8
	.723	15.7	9.6	7.0	4.0	3.7	-2.2
	.889	18.4	7.1	-6.5	-8.7	-9.6	-10.1
OZONE (PPM)	.853	.163	.037	.033	.012	.011	.003
	.755	.133	.100	.089	.080	.075	.002
	.797	.041	.040	.010	.002	.002	.002
CO (PPM)	.755	3.81	1.41	1.04	.16	.14	.13
	.996	43.5	.51	.24	.07	.06	.05
	.932	4.2	.42	.37	.13	.09	.07
CH_4 (PPM)	.789	4.2	1.8	1.7	1.3	1.0	.06
	.941	5.6	1.8	1.8	1.6	1.5	1.4
	.931	4.1	1.7	1.7	1.5	1.5	1.4
THC (PPM)	.743	3.10	1.97	1.97	1.63	1.58	1.40
	.734	4.19	1.91	1.78	1.28	.91	.05
	.924	4.31	1.80	1.59	1.50	1.36	1.25

[1/] RATIO ONLY USED TO FLAG HIGH VALUES.

TABLE 9

EXAMPLES OF FLAGGED VALUES FOR SEVERAL VARIABLES
USING DIXON RATIO ON 13 RURAL STATIONS IN THE RAPS DATA BANK[1]

VARIABLE	DIXON RATIO	STATION VALUES					
		HIGHEST	2 HI	3 HI	3 LO	2 LO	LOWEST
WINDSPEED (METERS/SEC.)	.815	6.4	3.9	3.8	3.3	3.2	3.1
	.808	4.1	2.2	1.5	.95	.84	.49
	.772	5.5	3.1	3.1	2.7	2.3	1.9
TEMPERATURE (°C)	.940	8.0	-3.1	-3.6	-4.3	-4.4	-4.9
	.957	8.5	-11.4	-11.5	-12.3	-12.4	-12.6
	.921	6.1	-7.0	-7.2	-8.3	-8.4	-8.5
DEW POINT	.845	31.4	6.7	.44	-5.1	-5.2	-7.8
	.782	18.1	5.0	4.6	2.1	.82	-2.9
	.701	33.2	17.6	17.3	13.4	10.6	6.9
OZONE (PPM)	.933	.206	.018	.016	.006	.002	.002
	.900	.252	.051	.048	.028	.026	.015
	.771	.045	.015	.012	.003	.002	.002
CO (PPM)	.929	8.7	.96	.77	.21	.17	.12
	.973	28.9	1.7	.94	.43	.17	.17
	.889	8.4	1.3	1.1	.28	.21	.14
CH_4 (PPM)	.944	10.1	2.4	2.3	1.9	1.8	1.7
	.873	7.6	3.9	2.7	2.1	2.0	1.9
	.791	5.0	2.9	2.7	2.1	2.1	2.0
THC (PPM)	.842	4.2	2.3	2.3	2.0	1.9	1.7
	.904	3.9	2.4	1.9	1.7	1.7	1.3
	.855	5.4	3.0	2.2	1.9	1.7	1.4

---

[1] RATIO ONLY USED TO FLAG HIGH VALUES.

CLUSTER ANALYSIS AS A DATA VALIDATION TECHNIQUE

by

Harold L. Crutcher

(Consultant)
35 Westall Avenue
Asheville, North Carolina 28804

CLUSTER ANALYSIS AS A DATA VALIDATION TECHNIQUE

H.L. Crutcher

INTRODUCTION

In any study the collection, processing, and storage of data are fundamental. Contaminated, adulterated, or "noisy" data confuse the investigator. Data do not necessarily fall into neat categories. Usually there are mixtures. Some of these are determinant; some are not.

There are many techniques used to cluster and to classify data. This paper discusses one. This technique separates mixed data sets into subsets. Each subset will exhibit homogeneous characteristics. The investigator can assess the relative importance of the subsets and the nature of the subsets. Outlying subsets may indicate anomalous true conditions or may indicate malfunctioning of some part of the observational program.* Thus some idea of data quality may be obtained.

The techniques used here require the assumption of the normality of distribution of the data. If the data are not normally distributed, then some transformation to approximate normality should be made. The logarithmic transformation is often used. Where this is known to be inapplicable, then another transformation is needed. For example, cloud cover is not well represented by the normal nor the log-normal distribution.

The clustering program discussed here was initially developed by Wolfe (1) and modified by Crutcher and Joiner (2). It will accept any input data. However, the criteria selected by the user to enable the computer to make decisions are based on the assumption of normality of distribution. Any departure from this assumption introduces some uncertainty in the results.

---

* In particular, the outlying subsets may be examined for their validity. The minimum number in a set is determined by the number of elements being examined simultaneously. With five elements, the minimum subset will be one more than five, or six.

There is always enough uncertainty without introducing more. For example, although the 0.05 probability level is selected for decision, departure from normality may actually cause the decision to be made at some other level, but this level will never be known.

ESSENTIAL PHILOSOPHY

Many elements and many observations may be treated. Computer capacity, time, and money will be the controlling factor. Within these constraints the user may wish to randomly select a representative sample of the data for processing.

Most investigators choose to standardize their data. This produces dimensionless numbers with means of zero and a variance of one for each element. A mean of zero and a variance equal to the square root of n, the number of variables, are obtained in the multivariate case.

If the elements are uncorrelated and are homogeneous, a spherical cluster of data points is obtained. If the data are correlated, the original element axes are rotated so that along the new axes obtained, the new components are not correlated. The new system will then be spherical in shape if the data are homogeneous.

CLUSTERING

If the data are clustered, even though the data are standardized, tests of normality will be rejected. Therefore, the usual procedure is to cluster data into probable groups. Then null hypotheses are established to compare two groups against one, three against two, and so on until the null hypothesis is not rejected.

Initial

The computer program may be set to establish any number of initial clusters. Here, the first 40 entry data serve to establish 40 clusters, but arbitrary clusters could have been inserted.

The number of elements is n so each datum is an n-vector with its point in n-space. The 40 clusters represent 40 centroidal points in n-space. The distances between the centroids are computed. The two closest are merged to a new centroid which is a mean or average of the two. After merging, there are now 39 clusters. A new datum enters from storage to again fill out the 40 spaces reserved. This procedure is repeated over and over again until all data have entered and have been assigned to one of the clusters. Variance considerations or other distance measurements as well as distances between the centroids can be used.

Intermediate

Forty clusters were obtained initially. These forty clusters are compared on an argument of the distance between centroids or on variance considerations as before. The two most nearly alike are merged into one cluster. The procedure continues until one final cluster remains.

Figure 1, which came from Figure 1 Crutcher and Joiner (3), illustrates in an abbreviated way the flow of 74 observations until final coalescence into the final group. The final group is made up of the initial data but the sequence is altered to show the entrance into the group. The data are 4-space upper air observations at the Canton Island 30-mb surface (1960-1964). The four elements of the subspace at the 30-mb surface are:

1. Height of the surface;
2. Temperature of the surface;
3. East-west wind component;
4. North-south wind component (orthogonal to east-west component).

Final

The user is required to provide the number of clusters wanted for review and the probability level of rejection for the null hypothesis. The sequential tests are for k+1 groups versus k groups where k runs from 1 to 40. The tests will continue until the null hypothesis in not rejected or until the requested number of clusters have been examined.

Output

At the completion of the computer program, output is presented as:

1. The initial set of data in some sequence established by the use.
2. Matrix of observational setup and preliminary comparisons.
3. Forty groups (clusters) with identified input data and means.
4. Coalescence, step by step, into fewer and fewer groups until the unlike final group is obtained.
5. Statistics with means and standard deviations for the main group

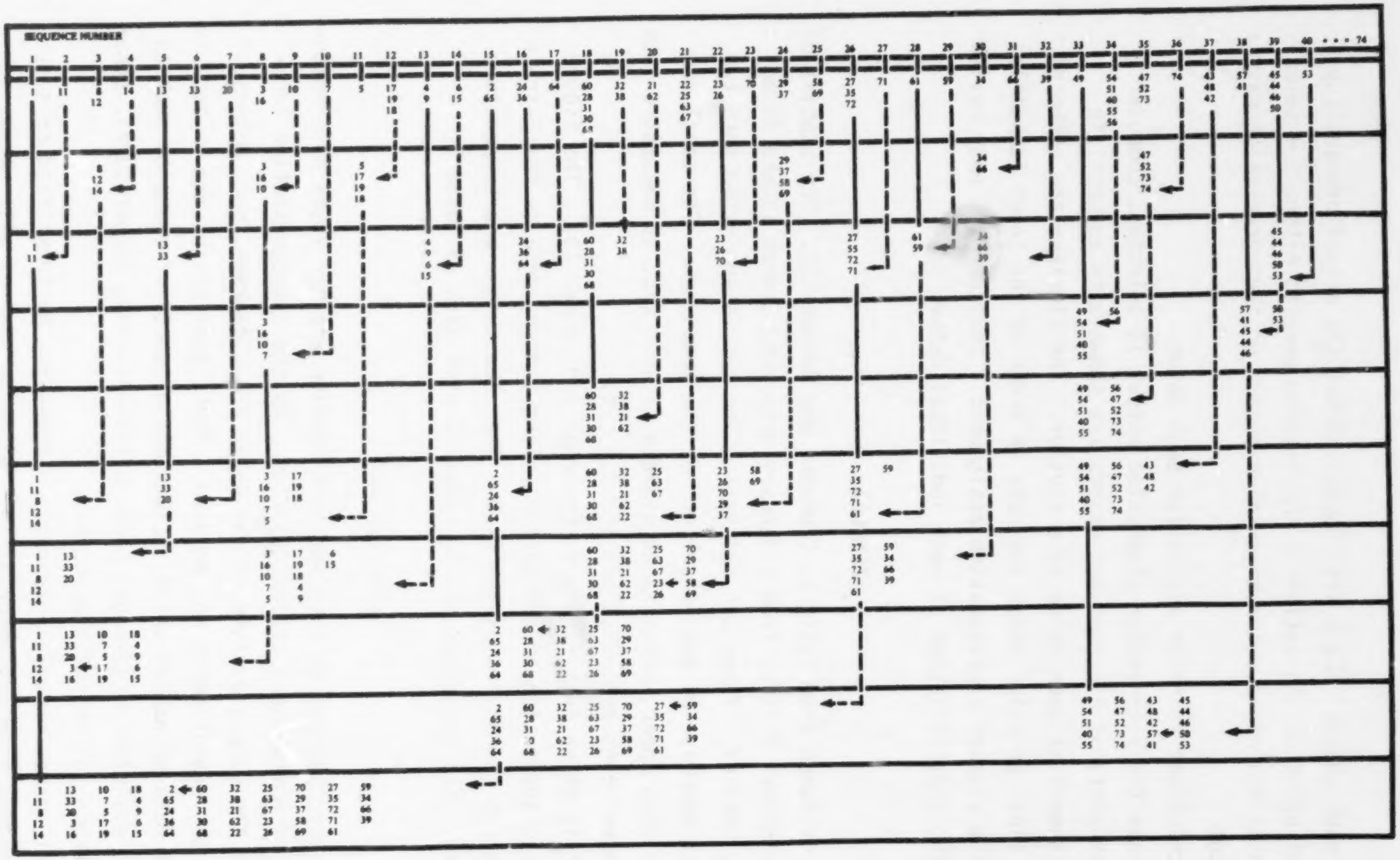

SEQUENCE NUMBER
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 . . . 74

FIGURE 1. SCHEMATIC CLUSTERING DIAGRAM FOR 74 FOUR-COMPONENT VECTORS OF CANTON ISLAND, 30-MB HEIGHT, TEMPERATURE, ZONAL COMPONENTS, AND MERIDIONAL COMPONENTS OF THE WIND (1960-64)

and each subset. Each set comparison has the actual probability level printed. An option in the program permits either the selection of the eigenvector-eigenvalue output or the correlation matrix output.

6. Discriminant function scores for each datum.
7. In case the eigenvector-eigenvalue output is selected, computer print-plots of discriminant scores are shown with appropriate assignment of each datum to a cluster. The clusters are numbered.
8. The final printing shows the data in order of the input identified by the cluster configuration assignment. This permits easy review of the classification of each individual datum.

EXAMPLES

Table 1 is taken from Table 5, Crutcher and Joiner (3). The statistics are for two clusters derived from a January data set, Canton Island 30 mb surface data, 1953-67. There are 434 data. The set is separated into two clusters which comprise 31 and 69 percent of the total set. The zonal component of the wind speed, which is an average of -6.4 m/s, is separated into two groups whose means are -28.1 m/s and 3.5 m/s.

Figure 2 is taken from Figure 3, Crutcher and Joiner (3). The figure exemplifies the separation of the Canton Island January data in the 2-space of the orthogonal components of the wind. The mean height and temperature data are shown. The variances may be compared with data of Table 1.

APPLICATIONS

Data of any type may be examined to determine whether there are reasonable subsets of homogeneous characteristics. As the standardization techniques remove the dimensionality of the data, i.e., degrees, mph, meters, grams, etc., any measurements may be used. Thus, application may be made to environmental data ensembles which include measurements of elements such as particulates, pollutants (gaseous), precipitation, wind, temperatures, pressures, or changes of any of the above.

Pollutant source or likely deposition areas may be identified or suggested. Extension from one observational point to several will permit

TABLE 1. CANTON ISLAND, 30-MB DATA AND THEIR SEPARATION INTO TWO CLUSTERS, 1953-67

		January $N = 434$			*April* $N = 509$	
	Group 1 (total)	Group 2 (easterly)	Group 3 (westerly)	Group 1 (total)	Group 2 (easterly)	Group 3 (westerly)
Data fraction	1.000	0.310	0.690	1.000	0.427	0.573
$\bar{H}$ (gpm)	23764.9	23720.6	23785.0	23801.3	23773.8	23821.9
s_h (gpm)	94.9	65.6	100.0	81.8	71.0	84.5
$\bar{T}$ (°C)	−57.0	−57.4	−56.9	−54.9	−56.1	−54.0
s_t (°C)	2.7	2.3	2.9	3.0	2.7	3.0
$\bar{u}$ (m s^{-1})	− 6.4	−28.1	3.5	− 4.6	−20.9	7.6
s_u (m s^{-1})	16.2	5.8	7.5	16.7	10.7	7.6
$\bar{v}$ (m s^{-1})	0.4	1.1	0.0	− 0.0	− 0.2	0.1
s_v (m s^{-1})	4.3	4.7	4.1	4.0	3.7	4.2
r_{uv}	− 0.1	0.2	0.1	0.0	0.0	0.0
		July $N = 558$			*October* $N = 476$	
Data fraction	1.000	0.470	0.530	1.000	0.330	0.670
$\bar{H}$ (gpm)	23938.0	23879.6	23989.7	23886.6	23829.3	23915.0
s_h (gpm)	94.0	78.5	75.6	104.4	67.6	108.4
$\bar{T}$ (°C)	−53.6	−55.4	−52.1	−55.1	−55.9	−54.7
s_t (°C)	3.1	2.2	3.0	2.8	2.2	2.9
$\bar{u}$ (m s^{-1})	− 4.3	−23.4	12.6	− 3.8	−30.2	9.3
s_u (m s^{-1})	19.6	5.8	9.3	19.7	4.7	7.3
$\bar{v}$ (m s^{-1})	0.1	0.3	− 0.2	0.2	0.0	0.3
s_v (m s^{-1})	3.9	3.3	4.3	4.2	3.7	4.4
r_{uv}	− 0.0	0.1	0.1	0.0	− 0.1	0.0

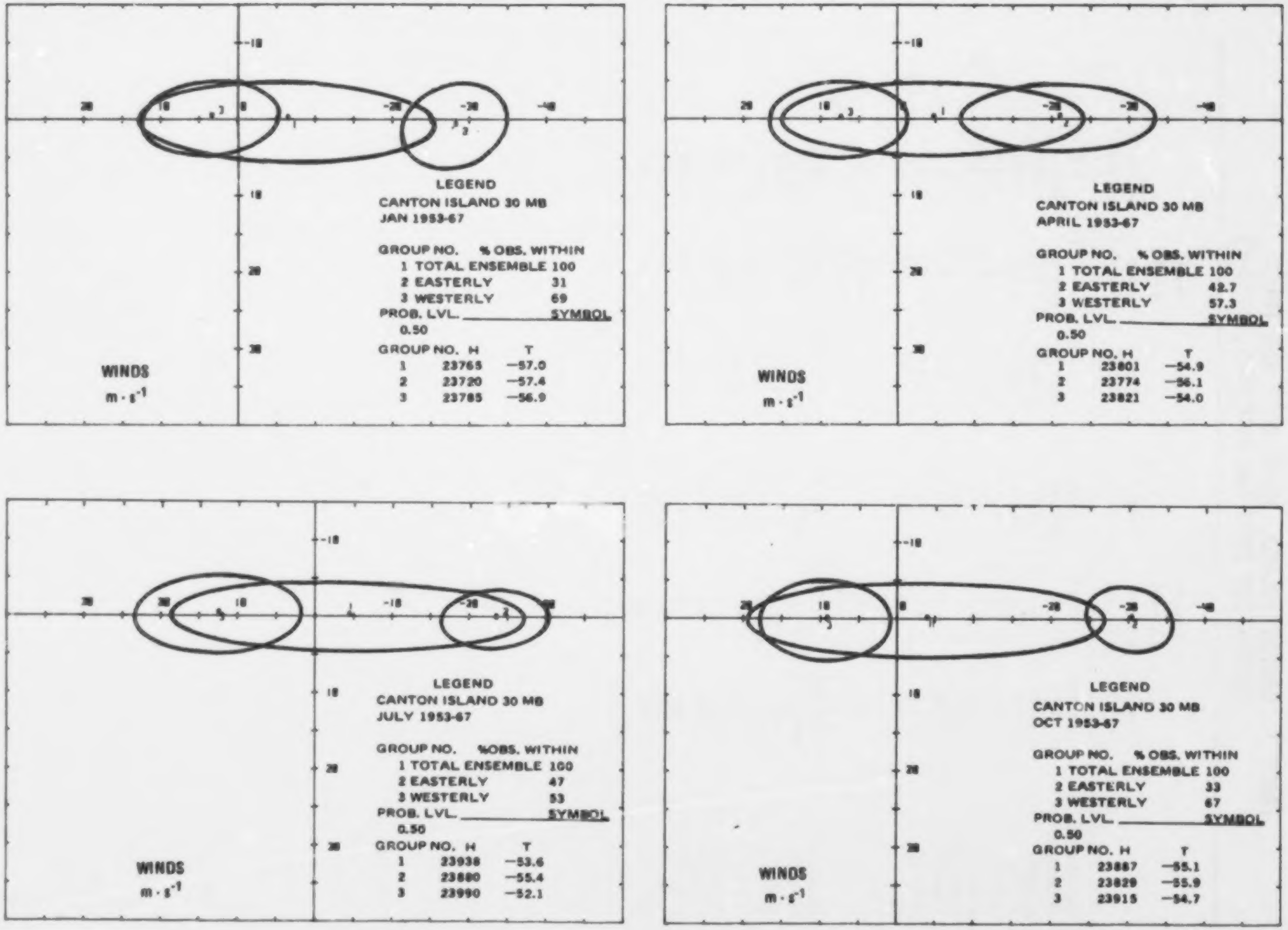


FIGURE 2. GEOMETRY OF THE WIND DATA SHOWN IN TABLE 1

mapping in a topographical sense. These techniques can be applied to data sets derived from topographical studies made by use of polynomials, trigonometric, polynomial orthogonal, or other types of polynomials. Recent computing advances in the computation of orthogonal polynomials known as asymmetric singular decomposition (ASD) procedures make these polynomials easier to obtain and use.

REFERENCES

1. Wolfe, J.H. NORMIX 360 Computer Program. Research Memorandum SRM 72-4, Naval Personnel and Training Research Laboratory, San Diego, CA (1971) 125 pp.
2. Crutcher, H.L. and Joiner, R.L. Separation of Mixed Data Sets into Homogeneous Sets. NOAA Technical Report EDS 19, National Oceanic and Atmospheric Administration, Asheville, North Carolina 28804 (1977) 165 pp.
3. Crutcher, H.L. and Joiner, R.L. Another Look at the Upper Winds of the Tropics. J. Applied Meteorology, 16 (5), (May 1977) pp462-476.

ENGINEERING COMPUTATIONS AND DATA COLLECTION FORMATS USEFUL IN DATA VALIDATION

by

A. Carl Nelson, Jr.

PEDCo Environmental, Incorporated
505 South Duke Street
Durham, North Carolina 27701

ENGINEERING COMPUTATIONS AND DATA COLLECTION FORMATS USEFUL IN DATA VALIDATION

A.C. Nelson, Jr.

A considerable number of "after-the-fact" data validation techniques will be given during this one-day conference. It is not the attempt here to try to summarize all of these techniques but to indicate some of the important validation procedures from the view point of the laboratory and field experts. The approach taken is as follows. The question was asked of laboratory and field experts: What are some of the important areas of data validation in order to yield data of good quality? Some of the areas listed are briefly described below for both ambient air monitoring and source testing.

Data Validation - Ambient Air Monitoring Data

1. Audits

The purpose of audits should not be to point a finger at the organization/team being audited. In some cases the auditor can be in error. The value of an audit is that it can identify a gross bias or inaccuracy in reported data. One recent example of a problem was in the use of an incorrect method. The audit pointed out the problem, special instruction was given, and the condition was hopefully corrected. It is possible that no after-the-fact data validation techniques could have identified the error in this case since there was a bias throughout the region as the same procedure was taught to all operators. The audit has served its purpose well in this example.

In EPA sponsored audits the auditor is almost always checked out at EPA, Quality Assurance Branch (QAB) prior to conducting a field audit. This provides a traceability to a common standard and method.

2. Knowledge of the instrument

This is certainly one of the most important considerations in obtaining good quality data. The operator must know the sensitivities/interferences of the instrument. An example would be the interference of CO_2 concentration on an SO_2 analyzer employing a Flame Photometric Detector. A test was designed to test the possible sensitivity and the results indicated a definite and reproducible relationship.

This particular type of error detected for a particular analyzer could be very difficult to detect by after-the-fact validation procedures. Some independent and accurate check must be made using an instrument which has been tested for possible interferences/sensitivities. If the sensitivity of an instrument to an interference has been precisely determined and is reproducible, then a correction can be made to obtain the result which would have been observed if no sensitivity existed, this was true in the case mentioned.

Another means of gaining information about the instrument/method is to design a ruggedness test to check out possible gross factors/steps which may have a significant effect on the results or measurements if appropriate control is not exercised.

3. Interlaboratory tests

Participation in these tests provides a means of checking the laboratory analysis methods and of validating the current multipoint calibration curve. The feedback of information from the laboratory performing the overall analysis of the results from all participating laboratories is most important. For example, if a laboratory is consistently in error for a particular analysis or range of concentrations, then this laboratory must have some means of correcting this problem through communication with the overall test laboratory or some representative thereof.

The performance survey is a very good means of validating data. Furthermore, it is also a good source of information about what can be expected from a particular analysis procedure.

In a conversation with a supervisor in one laboratory, he indicated that they were performing their CO analysis incorrectly and that the performance survey helped them to identify a problem which they did not know they had.

On a small scale, three or four laboratories could set up their own interlab test for a particular analysis for which no performance survey data are being obtained by EPA, NIOSH or some other agency.

4. Standards traceable to an NBS standard

Often one hears that calibration gases are not accurately analyzed. It is thus necessary that the user check the calibration gases prior to their use in developing new calibration curves. All measurements must be traceable ultimately to a primary standard.

Significant errors in calibration gases can usually be determined by a check against the previous calibration curve obtained using the most recent gas. EPA, QAB has developed a protocol for traceablity of gases.

5. Data reduction

The raw data must be recorded legibly and completely on appropriate data formats. The calculations should be checked either completely or on a sampling basis. In this manner the equations used, the substitution of the correct values in these equations, and the calculated results are all checked. This should be an internal audit as well as a part of an external audit.

6. Other considerations

Some other considerations are the more routine types of quality control and assurance techniques which are primarily internal functions. Some of these techniques are the use of blind reference samples, quality control limits for internal checks of reference samples, comparison checks of two or more calibrators, ruggedness tests, internal audits by an independent operator, and chain of custody procedures.

Data Validation - Source Tests

The results of the first series of collaborative source tests clearly showed that more quality control and data validations were needed to ensure good quality data. The results of the first Method 5, particulate collaborative test, using average testing teams and no special quality control, produced a relative standard deviation for each run in excess of 50%, with the outliers thrown out. As a result of this poor reproducibility, several quality control and data checks were incorporated into the collaborative test series. These controls and the use of selected testing firms produced results that were repeatable to within about 10 percent for each run. Most of these additional quality control checks are now detailed or implied in the revised methods contained in the Federal Register, August 18, 1977. The collaborative test series showed two other areas of concern with respect to quality control and data validation. The first was that the methods and additional written procedures were thought to be clearly written as to their execution. This was found not to be the case in the early collaborative tests as many variations were noted in the performance of the methods. It became obvious early in the program that the performance of the average testing team should be observed by a qualified observer. Also the nature of most of the errors were such that they could not or would not have been detected by any data validations on the emission test report.

The second area of concern was that most of the quality control techniques were executed prior to the performance of the field test and the assumption was made that all components remained unchanged during testing. Two examples are dry gas meter calibration and the pretest leak check. If the dry gas meter calibration changed during testing or if a leak developed in the sampling train, this would not be detected.

The collaborative testing program has clearly demonstrated that to properly perform the needed data validations, controls must be clearly defined and observed before, during and after each test series. Data validation of uncontrolled and unobserved sampling is not effective as a general rule and usually will not clearly determine acceptability or unacceptability of data.

The revised method published August 18, 1977 contains many equipment performance calibrations and validations. The examples used before, a dry gas meter calibration and leak check only prior to testing, have now been changed to include a post-test leak check and a post-test meter calibration.

The best method for data validation has become equipment performance validation. If the equipment is operating properly the data should be accurate/precise within the determined limits for that method.

Source Test, Report Review[1]

One aspect of source testing report review involves checking the results, not only that the correct equations were used and there were no mathematical errors, but also that the correct values were used as inputs into the equations. The latter requirement can be checked quickly if all of the required data were measured and recorded legibly, and the raw data sheets are submitted with the report. Any report which includes a computer listing of the raw data, instead of the original data sheets, should be rejected.

The degree to which the calculations should be checked is generally a function of the consistency of the results and the reviewer's confidence in the tester's ability. The various levels of review possible for the calculations would be (1) none at all, (2) random spot checks, (3) complete review of results which seem inconsistent, with respect to each other or to typical results, (4) complete review of one randomly chosen run, and (5) complete review of all runs.

There are some empirical techniques that can be used to check or validate process and sampling data provided by the tester and the source. In some cases, the sampling data from the tester can be used to check process data supplied by the source. Some of the available techniques are given herein. The experienced reviewer will ultimately develop his own list of short cuts, cross checks, and rules of thumb.

---

[1] DSSE Workshop - draft report.

1. Barometeric Pressure

Incorrect barometric pressure measurement will not generally cause errors of more than 10 to 15 percent, but it is a very common error. The value reported by the tester can be checked in two separate ways: (1) At sea level, the barometric pressure is almost always between 29 and 31 inches of mercury, and usually close to 30. For every 1000 feet above sea level, the value will decrease by 1.1 in. Hg. Therefore, if a test is run in Denver, with an elevation of 5000 feet above sea level, the barometric pressure reported should be from 23.5 to 25.5 inches of mercury. (2) The reviewer can call the airport closest to test site, and ask for the "station" pressure (not corrected to sea level) for the date of the test.

2. Leak Tests

If the report claims that leak tests were performed either before each test or after filter changes, the dry gas meter readings on the data sheet would indicate this. In other words, it is unlikely that a leak test was done before run #2 if the final volume reading for run #1 is the same as the initial volume reading on run #2. If a leak test was made in the middle of the run (because of a filter change, for example), the volume readings before and after the leak test would be shown on the data sheet, so that the computed meter volume could be adjusted accordingly.

3. Moisture Data

The results presented in the report for the volume percent of water vapor in the gases sampled can be checked in several different ways. For any combustion source, the moisture content can be approximated by use of nomographs[1] if the reviewer calculates the excess air and can estimate the ambient temperature, ambient humidity, and the free water in the fuel. Hopefully, the process data will include an analysis of the fuel. If not, use zero for gas and oil, 10 percent for bituminous coal, and 25 percent for lignite, bark, wood, and refuse unless the fuel has been rained on recently. If the best estimates available are ranges, use the high and low estimates to bracket the moisture content.

Entrained droplets of liquid water in the stack gases can yield an erroneously high moisture content. All moisture data should be checked (even if there are no entrained water droplets) to ensure that the reported value is not higher than the saturation moisture content. Nomographs[1] provide moisture content at saturation as a function of stack absolute pressure and stack gas temperature. If the reported value is higher than the maximum read from the nomograph, the data are suspect. Generally, if the high reading was caused by entrained water droplets, the value is adjusted to the saturation moisture content.

---

[1] DSSE Workshop - draft report.

In sources where the process involves drying (removing water) from a raw material or product, a water balance across the process should validate the moisture data in the report. Remember to include the water introduced as humidity in the ambient air.

4. Orsat Data

For any combustion source, the relative amounts of oxygen and carbon dioxide in the flue gases can be predicted by the use of a nomograph.[1] When a report is submitted containing orsat data (or CO_2 and O_2 data from any other instrument), the data can be checked by aligning the type of fuel with the $\%CO_2$, and checking the $\%O_2$ from the nomograph with the reported value. If the results do not check, it indicates that there is a problem with the reported data.

This nomograph also gives the percent excess air based on the type of fuel and orsat analysis. The reviewer should be cautioned that if the orsat data were taken after a water scrubber, the nomograph will not work, since the scrubber will remove an indeterminate amount of carbon dioxide.

5. Volumetric Flow Rate Data

The volumetric flow rate is difficult to cross-check accurately, but there are several ways of determining if the reported values are in the "ball-park". In any duct or stack where the air is moved by a blower, the design criteria generally result in a gas velocity of 25-40 feet per second.

---

[1] DSSE Workshop - draft report.

The idea is that higher velocities cause prohibitive pressure losses, and lower velocities are uneconomical due to the cost of the duct work. Since the size of many stacks is dependent on structural strength or future needs, the check works best for the duct work leading to the stack. If the velocity measurements are made in the stack, and the stack cross-sectional area is much larger than that of the duct work, apply the 25-40 feet per second check by dividing the volumetric flow rate by the duct area. If there is no fan or blower in the process, such as with a natural draft boiler or incinerator, the flow will generally be 5-15 feet per second. Keep in mind that the ranges given here are not theoretical limits, but merely commonly encountered values. If the test results presented do not fall within these ranges, it is only a signal to look at the velocity data more closely.

In reviewing test results, it is always desirable to have available the results from any previous tests on the same source, previous tests on any similar sources (such as an identical unit at the same plant), or tests performed at the inlet to the control device. If the inlet tests were done simultaneously with the outlet tests, the volumetric flow rates (corrected to standard conditions) should match from inlet to outlet. If the control device uses water, the checks should be made on a dry basis. Air leakage in or out of the control device can occur, which would lessen the value of this check, but air leakage can generally be identified by a change in the moisture, temperature, or CO_2 content from inlet to outlet. Since inlet tests are not usually done for compliance, they are often performed in ducts with little or no straight run, which can cause higher than real velocity data, a factor to consider when making inlet-outlet comparisons.

Many sources have fan performance curves for the fans used in the process, and these can be used as a check against the reported flow rate data. The gas flow moved by the fan is a function of the pressure head produced (or induced, or both), the gas temperature, the gas composition, and the fan speed (rpm). Unless all of these factors are controlled or quantified (which is a rare situation) the fan curves can only be used to estimate or roughly check the flow rates.

When process equipment and/or control devices are designed, there is generally a design specification on volumetric flow rate. If these specifications are available (from the source or from permit forms) they can be used to check the tester's results.

6. Process Data

There are probably as many different ways to check process data as there are types of processes. Some are so much a part of a particular process that they could not all be discussed here. Therefore, if the checks mentioned herein are not adequate for the process in question, then that process should be studied (using the literature and communications with the source) to determine if some additional checks are available for use.

For many processes, the production rate is relatively constant from day to day. In this case, the production rate reported should compare favorably with the annual production rate (or annual raw material usage rate) divided by the number of operating days.

In a case where the reviewer wants to compute the production rate from the raw material rate, or compute the raw material rate to check the production rate, the principle of material balances should be employed. If one ignores nuclear reactions, then it can be stated that in any process, matter will be neither destroyed nor created. This means that any materials entering the process must either accumulate or leave the process (in minus out equals accumulation). The material balance can be done on all components of the process stream, or it can be limited to a single component such as water or carbon dioxide.

Drying operations, like grain dryers, are good examples of sources which adapt readily to a water balance. Water enters the process from the grain itself, from the drying air (which is generally ambient air), and from the combustion of fuels containing hydrogen; it leaves as water vapor in the exhaust and as residual water in the grain. The stack test data provide the total water vapor leaving the dryer, by multiplying the total gas flow rate by the percent water vapor, and converting the result to a mass rate. From the amount of fuel burned, one can compute the water vapor produced by the combustion. If the ambient temperature and relative humidity are known, the water supplied by the drying air can be computed. From the symbols shown in Figure 1, the

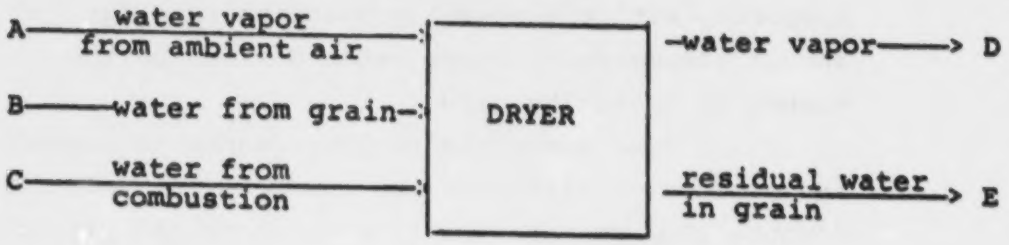

A
water vapor from ambient air
B
water from grain
C
water from combustion
DRYER
water vapor
D
residual water in grain
E

Figure 1. MATERIAL BALANCE

water balance would be:

$$A + B + C = D + E,$$

and A, C, and D have been computed. If F is the weight of grain dried, W is the inlet moisture fraction for the grain, and W' is the outlet moisture fraction for the grain, then

$$B = WF, \text{ and } E = W'F.$$

Substituting these expressions in the above equation and solving for F yields:

$$F = \frac{D-A-C}{W-W'} .$$

7. Emission Results

Unfortunately, the most difficult data to validate are the emission results, which also are the most important data to validate. Emission rates for gaseous pollutants, such as SO_2 and NO_x, can often be checked against process parameters, but this is because these pollutants are rarely controlled. For example, since essentially all the sulfur present in coal or oil will be liberated as SO_2 during combustion, a sulfur balance should yield a good check of SO_2 emission results. For a specific design of boiler or incinerator, the amount of NO_x produced can be estimated from the emission factor, published by EPA.[2]

The generation of particulate pollutants is a function of a large number of process parameters, many of which cannot be measured. Emission factors are available for many sources of particulates, but most of these sources use some type of control device to remove the bulk of particulates prior to the stack exhaust. As an example, consider the particulate emissions from a utility boiler with an electrostatic precipitator, or from an asphalt batch with a baghouse collector. The emission factor for these sources, prior to the control device, are listed in the emission factor book and the literature, and for now it will be assumed that they are accurate. The control devices that are used would have a design efficiency of 99 to 99.5 percent, but the actual efficiency could range from 50 to 99.9 percent. What this is saying is that if the uncontrolled emissions are 100 pounds per hour, the design efficiency would yield emissions of 0.5 to 1.0 pounds per hour, but the actual emissions could range from 0.1 to 50 pounds per hour. The emission factor book lists factors for various types of control devices, but these are design efficiencies, and the reviewer should resist paying much attention to them. They only reflect the emissions if the control equipment is operating at its design efficiency, and if that is assumed to be true, then there is no need to perform a compliance test in the first place.

---

[2] "Compilation of Air Pollution Emission Factors", U.S. EPA, Publication No. AP-42.

One approach which is often suggested (and used) by control agencies is the idea of comparing the three runs to one another. In other words, the validity of the data can be measured by the proximity of the three results to the average. This would work if all of the variation in the results was a function of random sampling errors. Using these assumptions data could be handled as in the following examples: (1) the three results are 2, 3, and 4, and the reported emission rate is 3, and (2) the three results are 2, 4, and 15, the 15 value is thrown out as an outlier, and 3 is reported as the emission rate. The second example says that since 2 and 4 are close to one another, and 15 is not close to 2 or 4, that the 15 must represent a gross sampling error, and only the 2 and 4 should be averaged together to get the emission rate.

Several additional considerations should discourage the reviewer from applying this validation technique. There is no question that three nearly identical results will instill confidence in the reviewer's mind, and that three widely different results will reduce that confidence. Using the example above, however, with the results of 2, 4, and 15, how can the observer tell what the rest of the "population" looks like? Had four samples been taken instead of three, with results of 2, 4, 15, and 15 and the last three were reported instead of the first three, the 4 would be thrown out as an outlier and 15 would have been reported as the emission rate. Process variations can occur during testing that could produce ten-to-one variations in the actual emission rates, and these variations can occur at any time, without any warning, and often without being noticed.

As a final note, any statistician can supply dozens of methods for evaluating a set of results, including ways to calculate confidence limits and eliminate outliers. Any statistician will also tell you, however, that a single set of three results is really too small to study statistically. And all the statistics in the world cannot replace common sense.

Summary

In summary, data validation must be an integral part of the data collection, analysis, reduction, and reporting process. Several useful data validation techniques which will aid in detecting large inaccuracies in the reported results are described. However, there are obvious limitations to the types of data inaccuracies which can be identified in both ambient and source tests as pointed out in this paper. It is hoped that the techniques suggested herein, from the viewpoint of laboratory and field experts, will stimulate further discussion.

Acknowledgement

Lawrence Elfers and William DeWees of PEDCo Environmental provided much of the information which is briefly summarized herein. "In addition, appreciation is due to Entropy Environmentalists, Incorporated because the draft copy of the Division of Stationary Source Enforcement Workshop contains information provided by this organization." I hope that this paper does not misinterpret their written and verbal suggestions. These inputs are greatly appreciated.

VALIDATION PROCEDURES APPLIED TO IN-USE
MOTOR VEHICLE EMISSION DATA

by

Marcia E. Williams

Office of Mobile Source Air Pollution Control
U.S. Environmental Protection Agency
Ann Arbor, Michigan 48105

VALIDATION PROCEDURES APPLIED TO IN-USE MOTOR VEHICLE EMISSION DATA

M.E. Williams

ABSTRACT

One of the functions of the Office of Mobile Source Air Pollution Control is the collection and subsequent assessment of data on the emission performance of in-use vehicles. On an annual basis, over 4.5 million fields of data are collected. These data must be carefully validated before they are used by EPA, by other government agencies, and by private citizens. The current data editing procedures are designed to be fairly routine but quite complete. Systematic problems are eliminated with thorough laboratory facility checkouts, frequent calibration checks, and correlation programs with the EPA laboratory.

The data editing procedure is divided into two parts - manual editing of the large amount of supporting data forms and strip charts and computer editing of all data cards. This edit procedure has detected error rates of from 14 to 32 percent manually and from 5 to 50% in the computerized phase. Most of these errors are correctable and less than five percent of tests are invalidated. Although some of the errors can be found in either the computer or manual phase, many errors can only be detected in one of the two phases. To avoid needless effort, the phases are performed in series rather than in parallel.

Editing costs are less than two percent of the total program cost and the current edit program is estimated to achieve a final error rate of about one percent. Future changes to the editing procedure focus on reducing the EPA manpower requirements without sacrificing the current quality level. An effort will be undertaken to determine how much effect various types of errors have on the ultimate uses of the data so that the question of "How good do the data have to be?" can be factored into the design of data validation methodology.

BACKGROUND

The Office of Mobile Source Air Pollution Control is responsible for generating a data base on the in-use emission and fuel economy performance of all mobile sources. These data are used by many groups within EPA as well as by other Federal agencies, state and local governments, private industry, and private citizens. A fairly complete list of uses for emission factor data is given in Table 1. Different degrees of data accuracy are needed for different data applications. Most applications are concerned with having an accurate estimate of average emissions or fuel economy. However, those items in Table 1 which are notated with an asterisk require that the data on every vehicle be completely accurate. At this point in time, the data edit procedure is geared to ensure that all fields of data are correct.

Table 2 provides a list of typical ongoing test programs. On an annual basis, OMSAPC spends between 2.5 and 4.5 million contract dollars on characterizing the performance of in-use vehicles. Typical ongoing test programs are listed in Table 2. Each test program involves the procurement and subsequent testing of consumer owned vehicles. Vehicle owners are given incentives such as a U.S. savings bond, a loaner car, and a free tank of gas to participate in the EPA test program. Vehicles are then tested over a variety of different test sequences. In some cases, entire test sequences are repeated with vehicles in different states of tune or with ambient test conditions varied. Table 3 lists the types of variables which are collected for each vehicle test. For each vehicle test sequence, there are 150 to 600 pieces of information gathered. For each vehicle tested, there are from one to six test sequences performed. Thus, on an annual basis, approximately 4.5 million fields of data are collected and must be validated.

GENERAL APPROACH

The data validation procedure begins with the assumption that there is no systematic bias or error in the data. Systematic errors are prevented by the development of detailed test procedures, record-keeping procedures, and mandatory recording formats. Each contractor must undergo a rigorous facility check-out at the beginning and end of each test program. The check-out includes performance tests for all contractor personnel including equipment operators, drivers, test technicians, and data handlers. In addition, EPA personnel specify frequent calibration checks on all equipment, carry out reference gas and vehicle correlation testing against the EPA laboratory in Ann Arbor, and implement both announced and unannounced contractor inspections throughout the duration of each test program. Table 4 details the procedures which are implemented to prevent systematic bias.

All EPA contractors are required to perform data validation before submitting any data to EPA. EPA contracts specify that the contractor must use some form of computerized edit procedure in addition to a manual procedure. However, the exact contractor procedure is not specified. Since contractors are not paid for tests until EPA accepts the tests as valid, an incentive exists to submit correct data as soon as possible after the completion of a vehicle test.

The EPA edit procedure is diagramed in Figure 1. The manual and computer aspects of data validation are carried out in series to avoid needless manpower effort and to ensure that at no time will data files contain any data that are not validated. The manual edit procedure is performed first and many of the steps in that procedure are listed in Table 5. The manual procedure concentrates on strip chart data including the driving trace and the emission concentrations. Although there is a trend toward contractor computerization of these items, a fully computerized data aquisition system is expensive and is not required by EPA due to the short term (annual), fixed-price nature of EPA contracts. Thus, errors in following the appropriate driving trace and properly zeroing and calibrating analyzers can only be detected in the manual phase.

Table 6 presents the types of checks which are performed in the computer edit procedure. As shown in Figure 1, the computer editing does not occur until the manual edit checks indicate a potentially valid test. Table 7 summarizes the types and the severity of errors which have been detected. Tests are invalidated only in cases where test procedure errors are uncovered or in cases where key data are missing. Table 8 lists typical reasons that complete test sequences have been invalidated.

The one type of error which the current edit procedure is not specifically designed to detect is discrepancies between the computer cards and the supporting documentation. If each computer card entry is within range and consistent with other computer data fields, it will not be flagged. For example, if a highway emission result is incorrectly keypunched as 4.52 instead of 4.92, it would not be detected since both numbers could be equally valid. One would have to examine the analyzer trace and the data packet notation to know which value was correct. Since data are double keypunched, these types of errors are assumed to be minimal. However, consideration is being given to the implementation of an acceptance sampling procedure to ensure that information on the data cards matches information in the supporting documentation.

RESOURCES AND RESULTS

Table 9 indicates the EPA detected error rates in three recent test programs. In each case, the error rate is the percentage of vehicles with at least one detected error; some vehicles may have multiple errors. The range of detected error rates clearly indicates that not all contractors employ the same levels of quality control. However, despite the high detected error rates, less than five percent of total tests are invalidated; most errors can be corrected.

Table 10 summarizes the required EPA editing resources for two recent test programs. These resources are examined as a fraction of total contract cost in Table 11. Assuming that EPA manpower used to perform data editing costs $20,000 per manyear, the EPA cost for data validation is about two percent of total contract cost. With this resource effort, it is estimated that the undetected error rate is less than one percent.

FUTURE APPROACHES

Table 12 lists a number of additional data validation approaches. More automated data aquisition is being implemented in one ongoing contract. The system has taken considerable time to debug and it is too early to judge the cost-effectiveness of this approach. In recent contracts, EPA has increased the contractor data validation requirements. Again, it is too early to determine whether this action will prove to be a cost-effective way to achieve a low final error rate. Finally, in one large test program, EPA has stationed personnel at the contractor's site on a full time basis. Again, the effect on final error rate is not yet known.

Table 12 lists three additional approaches which have not yet been implemented. If improved contractor error rates can be achieved, EPA will attempt to reduce dedicated edit manpower by implementing a general spot check procedure. Such procedures are based on statistical principles and one such procedure is outlined in detail in Tables 13 and 14.

Before edit procedures can be implemented which attempt to lower the cost of editing by applying statistical procedures to key data fields, two important philisophical questions must be answered. First, it must be determined how good the data have to be. Variables of maximal interest need to be specified and in each case, the confidence and range within which the variable needs to be known must be determined. The second major area of uncertainty requires a determination of the impact that various errors make on the variables of maximal interest. Answers to these question areas are currently being pursued so that statistical edit procedures can be considered in more detail.

Table 1

USES FOR TEST DATA

EPA

1. ASSESSMENT OF EMISSION AND DETERIORATION RATES FOR AP-42 (HANDBOOK OF AIR POLLUTION EMISSION FACTORS)
2. DEVELOPMENT OF EMISSION AND FUEL ECONOMY CORRECTION FACTORS FOR AP-42
3. COMPARISON OF IN-USE LEVELS WITH CERTIFICATION, ASSEMBLY LINE LEVELS
4. DETERMINATION OF REASONS FOR POOR IN-USE VEHICLE PERFORMANCE
5. ASSESSMENT OF SHORT TEST/FTP CORRELATABILITY FOR ALT. 207(B) (APPLICABLE SECTION OF THE CLEAN AIR ACT)
6. ASSESSMENT OF INSPECTION/MAINTENANCE BENEFITS

*7. EVALUATION OF IN-USE VEHICLE COMPLIANCE WITH STANDARDS - SUPPORT FOR AGENCY RECALL PROGRAM

*8. COMPARISON OF PRODUCTION/PROTOTYPE FUEL ECONOMY LEVELS

9. SUPPORT FOR REGULATION DEVELOPMENT PACKAGES - ENVIRONMENTAL IMPACT ANALYSES
10. PRIORITIZATION OF AGENCY REGULATION/COMPLIANCE PROGRAMS

OTHER USERS

1. HIGHWAY ENVIRONMENTAL IMPACT STATEMENT (EIS) WORK
2. INDIRECT SOURCE REVIEW
3. REGION/STATE EMISSION INVENTORY WORK
4. EVALUATION OF IMPROVED PUBLIC TRANSPORTATION SYSTEMS
5. EVALUATION OF VEHICLE MILES TRAVELED (VMT) REDUCTION STRATEGIES
6. GENERAL TRANSPORTATION CONTROL PLAN (TCP) EVALUATION
7. FUEL AVAILABILITY STUDIES
8. AIR QUALITY MODEL INPUTS
9. STATE IMPLEMENTATION PLAN (SIP) CONFORMANCE WITH AMBIENT STANDARDS
10. HEALTH ASSESSMENT STUDIES

Table 2

TYPICAL ONGOING TEST PROGRAMS

1. ANNUAL IN-USE AUTOMOBILE TESTING PROGRAM
   - A. 7 CITIES
   - B. WIDE RANGE OF MODEL-YEARS
   - C. LARGE NUMBER OF EMISSION TEST CONDITIONS
   - D. @2000 VEHICLES PER YEAR
2. ANNUAL AUTOMOBILE RESTORATIVE MAINTENANCE TESTING PROGRAM
   - A. 3-4 CITIES
   - B. PRIMARILY NEW MODEL-YEAR VEHICLES
   - C. EXTENSIVE DIAGNOSTIC AND MAINTENANCE WORK PERFORMED
   - D. @400 VEHICLES PER YEAR
3. IN-USE LIGHT DUTY TRUCK TESTING PROGRAM
   - A. MULTIPLE CITIES
   - B. WIDE RANGE OF MODEL-YEARS
   - C. LARGE NUMBER OF EMISSION TEST CONDITIONS
   - D. @200 VEHICLES PER YEAR
4. IN-USE HEAVY DUTY TRUCK TESTING PROGRAM
   - A. SINGLE CITY
   - B. WIDE RANGE OF MODEL-YEARS
   - C. LARGE NUMBER OF EMISSION TEST CONDITIONS
   - D. @200 VEHICLES IN FY78
5. IN-USE MOTORCYCLE TESTING PROGRAM
   - A. 2 CITIES (HIGH AND LOW ALTITUDE)
   - B. WIDE RANGE OF MODEL-YEARS
   - C. LARGE NUMBER OF EMISSION TEST CONDITIONS
   - D. @250 VEHICLES IN CURRENT TEST PROGRAM
6. INSPECTION/MAINTENANCE DEMO PROJECT
   - A. PORTLAND, OREGON
   - B. 1972-1977 MODELS
   - C. LARGE NUMBER OF EMISSION TEST CONDITIONS AND AMBIENT CONDITIONS
   - D. @3000 VEHICLES AND 6000 TESTS
7. OTHER SMALL TESTING PROGRAMS
   - A. MOPEDS
   - B. DIAL-A-RIDE BUSES
   - C. AMBIENT TEMPERATURE TESTING
   - D. GOOD TECHNOLOGY VEHICLES

Table 3

TYPES OF VARIABLES COLLECTED FOR EACH TEST

1. Identification Data
   MODEL YEAR
   MAKE
   MODEL
   ENGINE DISPLACEMENT
   CARBURETOR VENTURIS
   CATALYTIC CONVERTER
   AIR PUMP
   NUMBER OF CYLINDERS
   TRANSMISSION TYPE
   VEHICLE IDENTIFICATION NUMBER (VIN)
   ENGINE FAMILY CODE

2. Emission Data
   IDLE DATA (2 POLLUTANTS)
   SHORT TEST DATA (3 POLLUTANTS, UP TO 5 TESTS)
   FTP DATA (4 POLLUTANTS, 3 BAG VALUES, COMPOSITE VALUE)
   OTHER CYCLES (4 POLLUTANTS)
   FUEL ECONOMY DATA
   EVAPORATIVE EMISSION DATA
   SULFATE EMISSION TESTS
   PARTICULATE EMISSION TESTS
   METHANE - NON-METHANE MEASUREMENTS
   MODAL TESTING

3. Ambient Conditions
   TEMPERATURE
   HUMIDITY
   BAROMETRIC PRESSURE

4. Test Conditions
   ROAD LOAD HORSEPOWER
   INERTIA WEIGHT
   SOAK TIME
   PRE-CONDITIONING SCHEDULE

5. Parametric Data
   ENGINE IDLE SPEED
   ENGINE TIMING
   MANUFACTURER SPEC VALUES
   COMPLETE DIAGNOSTIC CHECKS (9 MAJOR VEHICLE SYSTEMS, @10 COMPONENTS PER SYSTEM)
   TAMPERING DATA
   DRIVEABILITY DATA

6. Owner Questionnaire
   NUMBER OF TRIPS PER DAY
   NUMBER OF MILES PER YEAR
   TYPE OF DRIVING
   TYPICAL PASSENGER LOADING
   LAST MAINTENANCE PERFORMED - TYPE AND COST
   FUEL ECONOMY ESTIMATE
   TYPE OF GASOLINE USED

Table 4

PROCEDURES TO PREVENT SYSTEMATIC BIAS

1. SPECIFICATION OF GAS ANALYZER, DYNAMOMETER, AND CVS[1] EQUIPMENT (WITH PROVISIONS MADE FOR EQUIVALENT SUBSTITUTIONS).
2. EPA NAMES REFERENCE GASES.
3. EPA SPECIFIES ANALYZER CALIBRATION PROCEDURES INCLUDING GAS TYPES, CYLINDER FITTINGS, IMPURITY LEVELS, CURVE-FITTING PROCEDURE, AND REQUIRED ACCURACY.
4. SPAN GASES, SIMILAR TO CALIBRATION GASES, ARE SPECIFIED.
5. SYSTEM PLUMBING MATERIALS ARE SPECIFIED.
6. EQUIPMENT CHECKS ARE SPECIFIED
    - A. DAILY LEAK CHECKS - CVS AND ANALYTICAL SYSTEM
    - B. WEEKLY COMPLETE CURVE CHECK - ANALYTICAL SYSTEM
    - C. DYNAMOMETER WARM-UP PROCEDURES
    - D. COMPLETE CURVE CHECKS AFTER ANY SYSTEM MAINTENANCE
    - E. DYNAMOMETER CALIBRATED BI-WEEKLY
    - F. SAMPLE BAGS LEAK CHECKED BEFORE EACH TEST
    - G. DAILY NOx ANALYZER CONVERTER EFFICIENCY TEST
    - H. COMPLETE DAILY LOGS OF ALL GASES, CALIBRATIONS, MAINTENANCE, ETC.
7. MAXIMUM BACKGROUND LEVELS SPECIFIED.
8. COMPLETE EPA FACILITY CHECK-OUT INCLUDING TESTS TO CONTRACTOR PERSONNEL.
9. UNANNOUNCED EPA VISITS.
10. CORRELATION TESTING WITH EPA LAB.

---

[1] CONSTANT VOLUME SAMPLER

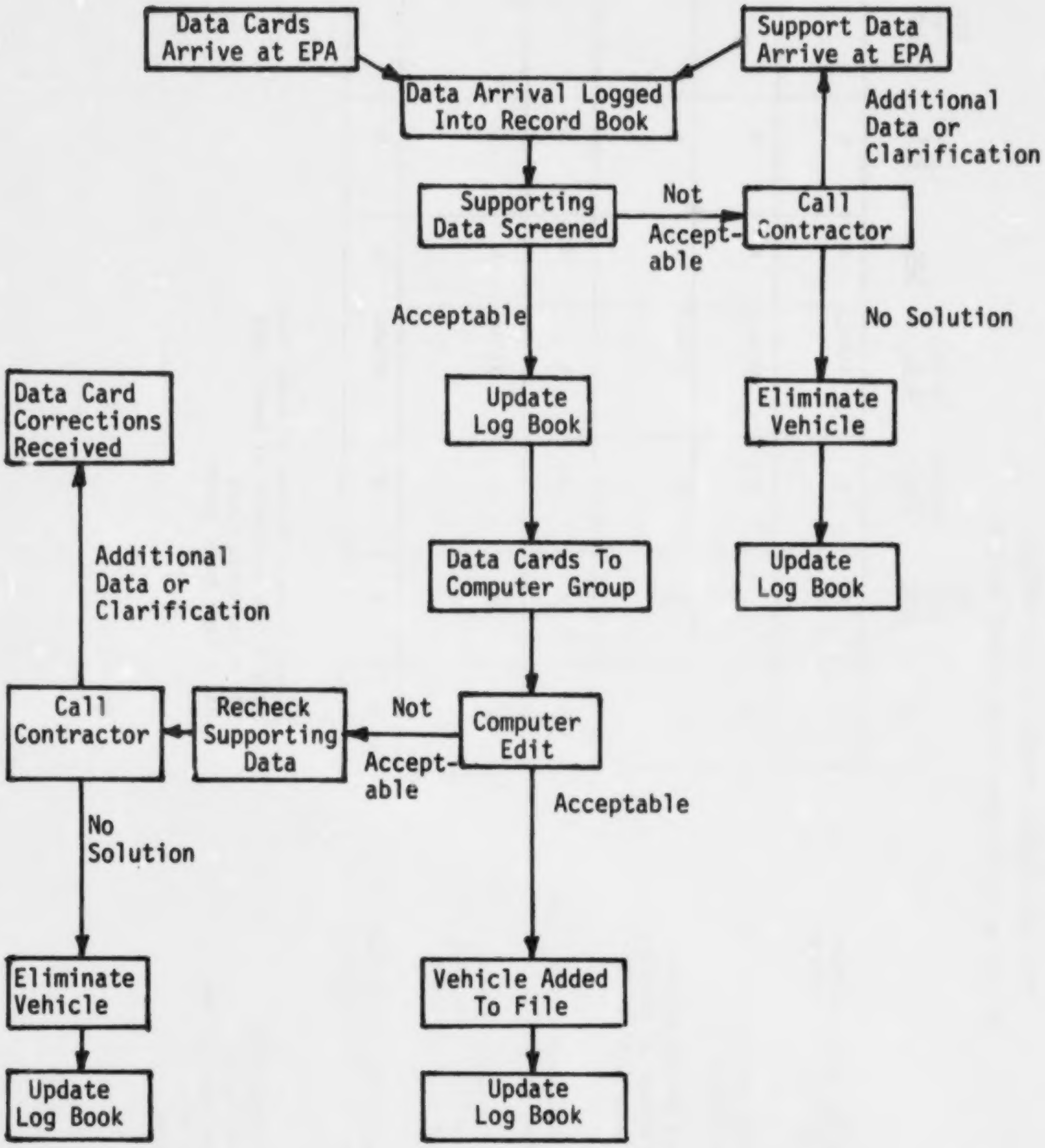

Data Cards Arrive at EPA
Support Data Arrive at EPA
Data Arrival Logged Into Record Book
Additional Data or Clarification
Supporting Data Screened
Not Accept-able
Call Contractor
Acceptable
No Solution
Data Card Corrections Received
Update Log Book
Eliminate Vehicle
Data Cards To Computer Group
Update Log Book
Additional Data or Clarification
Call Contractor
Recheck Supporting Data
Not Accept-able
Computer Edit
Acceptable
No Solution
Eliminate Vehicle
Vehicle Added To File
Update Log Book
Update Log Book

Figure 1. Flow Diagram of Edit Procedure

Table 5

FY74 LDV EMISSION FACTOR TESTING PROGRAM

OUTLINE OF MANUAL DATA PACKET REVIEW PROCESS

	FTP	FSC & HFET	Steady State	Surv. D.S.	LOS.	HIS.	FTM & CKM
Check analyzer strip charts for spans, zeroes, proper use of ranges, and correct readings of deflections.	X	X	X	X(bag)	X	X	
Make sure analyzer deflections are transferred correctly onto the data sheets.	X	X	X	X(bag)	X	X	
Make sure that the proper gas concentrations are entered on the data sheets by checking the values against the respective calibration charts.	X	X	X	X(bag)	X	X	
Check Barometer Reading on data sheets with barometer strip chart-when available.	X	X	X	X	X	X	
Check for proper reading of temperature test traces (wet & dry bulbs) and see that they are transferred correctly onto the data sheets.	X	X	X	X(bag)	X	X	
Make sure the CVS temp. trace is in specs. and appropriately entered on the data sheets.	X	X	X	X(bag)	X	X	

Key-
FTP - Federal Test Procedure
FSC - Federal Short Cycle
HFET - Highway Fuel Economy Test
CKM _ Clayton Key Mode
FTM - Federal 3 Mode

Steady State - Portion of modal test
SURV D.S. - remainder of modal test
LOS - Low Speed Cycle
HIS - High Speed Cycle

Table 5 (con't)

	FTP	FSC & HFET	STEADY STATE	SURV D.S.	LOS.	HIS.	FTM & CKM
Check to see that the temperatures recorded on the soak trace are within acceptable limits and make sure that the trace is for the proper time (i.e. approx. 12 hour period immediately **prior** to testing)	X				X	X	
Check Vehicle Check List 1. Scan vehicle info. (for proper recording of make, model, year, odom., eng.disp., carb., trans., & A/C)	X						
2. Check that old fuel was drained and correct amount of proper fuel was added.	X						
3. Check to see whether preliminary 10-minute drive was taken.	X				X	X	
4. Check to see that vehicle was taken out of soak in right time.	X				X	X	
5. Check to see what emission controls are present and transferred onto the mechanical inspection data sheet.	X						
6. Check measured & spec. values for idle eng. speed, dwell, timing, idle CO & HC. and see that they are entered properly on the mech. inspection data sheet.	X						

Table 5 (con't)

	FTP	FSC HFET	STEADY STATE	SURV D.S.	LOS.	HIS.	FTM & CKM
Check vehicle information data sheet							
1. space 7: proper site no.	X						
2. space 8: proper veh. no.	X						
3. space 12: proper match code	X						
4. spaces 13-16: proper test date	X						
5. space 21: proper test data	X						
6. space 27: proper make code	X						
7. space 29: use VIN to see what model was tested, and hence if proper inertial wt. and H.P. setting were used (space 60 & 64)	X						
8. space 42: proper model year	X						
9. space 44: proper make	X						
10. space 48: proper model code	X						
11. space 52: proper size code	X						
12. space 53: proper C.I.D.	X						
13. space 56: proper trans. code	X						
14. space 57: proper carb.	X						
15. space 58: No. of cyl.	X						
16. space 59: whether veh. had A/C	X						
17. space 68: whether A/C sim. was used and/or supposed to be used, and hence if proper add'n'l. H.P. loading was incorporated into test.	X						
18. space 69: check Odom. for reasonableness and correlation with O.Q.	X						
19. space 75: check fuel capacity to be sure proper amount of test fuel was used	X						
Check Driver's Trace							
1. Make sure driver's pen followed computer pen within specs.	X	X	X	X	X	X	X
2. Make sure crank times are not excessive.	X				X	X	
3. Make sure total time for cycle is reasonable	X						
4. Make sure any hold times are not excessive	X				X	X	

Table 5 (con't)

	FTP	FSC HFET	STEADY STATE	SURV D.S.	LOS.	HIS.	FTM & CKM
Check Owner Questionaire							
1. On basis of 1st 3 questions (new or used,) length of time since veh. purchase, & miles/yr. driven), determine if Odom. is reasonable.	X						
2. Check to see that amount of time vehicle is driven in various places for various purposes is reasonable (Q.# 4 & 5)	X						
3. Q's# 6-12, 14: make sure one and only one square is checked for each.	X						
4. On basis of whether or not veh. is equipped with cat. conv., check to see that Q# 13 (regarding Pb in fuel) is answered appropriately.	X						
Scan Preliminary Inspection Report	X						
Check to see that test agreement is signed by participant	X						
Check to see that veh. loan agreement is signed by participant	X						
Check to see that bond application card is present and filled out	X						
Scan vol./rev. value on data sheets for obvious error	X						
Check to see that appropriate No. of CVS. revs. are entered on data sheet	X						
Clayton key mode data sheet							
1. Check veh. wt. entered (space 19)						X	
2. Check for correct 1 HP (space 16)						X	
3. Check Hi Cruise speed (space 24)						X	
4. Check Lo Speedcruise (space 37)						X	
5. Check for reasonable correspondence between results for test 1 & test 2.						X	

Table 5 (con't)

	FTP	FSC HFET	STEADY STATE	SURV D.S.	LOS.	HIS.	FTM & CKM
Check computer print-outs							
1. Veh.#, date, HP, inertia	X				X	X	X
2. Baro, WB, DB, PIN, V/REV, TEMP, REVS	X	X	X	X(bag)	X	X	
3. All gas concentrations	X	X	X	X(bag)	X	X	X
4. Reasonable emission & MPG numbers	X	X	X	X(bag)	X	X	
5. Correspondence between bag values & running sample values				X			

Table 6

TYPES OF COMPUTER CHECKS

1. RANGE CHECKS ON ALL VARIABLES

   EXAMPLE: 68°F ≤ DRY BULB TEMP ≤ 86°F (VALID TEST RANGE)

2. ID CHECK FOR ALL TESTS WHICH SHOULD BE INCLUDED FOR EACH VEHICLE

   EXAMPLE: A SET OF 1/0 CODES ARE BUILT INTO THE VEHICLE ID INDICATING WHETHER FTP, HFET, EVAP., SULFATE, MODAL, SHORT TESTS ... ARE RUN. THEN, THE EDIT PROGRAM CHECKS FOR APPROPRIATE DATA CARDS.

3. ID INFO CHECKED AGAINST VIN

   EXAMPLE: MODEL YEAR CHECKED AGAINST VIN CODE.

4. FUNCTIONAL RELATIONSHIPS ARE DEVELOPED WHEREVER POSSIBLE

   EXAMPLES: MODEL YEAR RELATED TO MILEAGE; ROADLOAD HORSEPOWER RELATED TO INERTIA WEIGHT; ENGINE SETTINGS COMPARED WITH MANUFACTURER SPECIFICATIONS; ALLOWABLE EMISSION LEVELS DEPENDENT UPON MODEL YEAR; NUMBER OF CYLINDERS RELATED TO ENGINE DISPLACEMENT; FUEL ECONOMY RELATED TO ENGINE DISPLACEMENT.

5. COMPOSITE VALUES COMPUTED FROM COMPONENTS AND COMPARED TO CARD VALUE

   EXAMPLES: COMPOSITE FTP COMPUTED FROM INDIVIDUAL BAGS; FUEL ECONOMY COMPUTED FROM HC, CO, CO_2 DATA USING CARBON BALANCE.

6. RANKING COMPARISON OF RELATED VARIABLES

   EXAMPLES: IDLE MODE EMISSIONS LESS THAN HIGH SPEED MODE EMISSIONS; HIGHWAY FUEL ECONOMY GREATER THAN FTP FUEL ECONOMY; COLD EMISSIONS GREATER THAN STABILIZED EMISSIONS.

7. CHECK THAT EXPECTED BLANK COLUMNS ARE BLANK TO ENSURE PROPER COLUMN ALIGNMENT

8. DATA ON VEHICLE COMPARED TO PREVIOUS DATA ON SAME VEHICLE

   EXAMPLE: EMISSIONS TAKEN AT TWO DIFFERENT LOCATIONS OR TWO DIFFERENT TIMES ARE COMPARED.

Table 7

TYPES/SEVERITY OF DETECTED ERRORS

1. Errors in Test Procedure
   - A. Detected in manual and/or computer edit
   - B. Test is invalidated
   - C. Examples: driving trace out of specs (manual)
     wrong inertia weight setting (computer)
2. Errors in Calculation Methodology
   - A. Detected in manual and/or computer edit
   - B. All data are corrected by looking at packet
   - C. Examples: used wrong scale to read emissions (manual)
     composite FTP incorrectly calculated (computer)
3. Keypunch Errors
   - A. Detected in computer edit
   - B. All data are corrected by looking at packet
   - C. Example: engine displacement disagrees with VIN
     and/or is out of range
4. Missing Data
   - A. Detected in manual and/or computer edit
   - B. Test invalidated unless missing data can be found
   - C. Examples: driving trace missing from packet (manual)
     blank field for engine displacement (computer)
5. Discrepancy Between Data Card and Data Packet
   - A. Detected in computer edit check-out phase
   - B. Packet value assumed correct
   - C. Example: RLHP reading on data card is out of range
     and disagrees with what is recorded in the packet

Table 8

TYPICAL REASONS TESTS HAVE BEEN REJECTED

1. WRONG CVS COUNTS, EITHER TOO HIGH OR TOO LOW
2. EXCESSIVE CRANKING TIME, OVER 10 SECONDS WITHOUT REGARD FOR PRESCRIBED PROCEDURES FOR RESTART
3. WRONG INERTIA WEIGHT SETTING ON DYNAMOMETER
4. WRONG HORSEPOWER SETTING ON DYNAMOMETER
5. EMISSIONS CONCENTRATIONS READ OFF-SCALE OF ANALYTICAL EQUIPMENT
6. LABORATORY BACKGROUND EMISSION LEVELS TOO HIGH
7. VEHICLE HAS WRONG AXLE RATIO
8. SAMPLE BAGS NOT ANALYZED WITHIN 10 MINUTES OF TEST COMPLETION
9. DRIVER'S TRACE NOT FOLLOWED AS PRESCRIBED
10. RECORDING MALFUNCTION, 110°F DURING TEST
11. INITIAL FUEL TEMP. TOO HIGH (63°F), OR HIGHER
12. SOAK AREA TEMPERATURE TOO HIGH, FOR PRESCRIBED PORTION OF VEHICLE SOAK PERIOD
13. TEST AREA TEMPERATURE TOO HIGH FOR VEHICLE TEST PERIOD
14. CVS TEMPERATURE NOT WITHIN ±10° OF SET POINT
15. ANALYTICAL INSTRUMENT(S) SPANNED INCORRECTLY
16. TEST ITEM(S) NOT DOCUMENTED AS REQUIRED
17. ENGINE TIMING NOT CHECKED
18. ENGINE TIMING SET INCORRECTLY
19. ENGINE IDLE CO NOT CHECKED
20. ENGINE IDLE CO SET INCORRECTLY
21. ENGINE IDLE RPM SET INCORRECTLY

Table 9

CURRENT DETECTED ERROR RATES*

PROGRAM	MANUAL CONTRACTOR 1	CONTRACTOR 2	CONTRACTOR 3
1	14%		
2	17%	21%	
3	17%	26%	32%

COMPUTER (PROGRAM 1)

CONTRACTOR 1	CONTRACTOR 2	CONTRACTOR 3
5%	10%	50%

* PERCENTAGE OF VEHICLES WITH AT LEAST ONE ERROR DETECTED. LESS THAN 5% OF TESTS ARE INVALIDATED - MOST ERRORS CAN BE CORRECTED.

Table 10

Editing Effort Per Car (Manhours)

	Program 1	Program 2
LOGGING, FILING, SCOREKEEPING	.1	.1
INITIAL REVIEW	.3	.7
REVIEW AND SCOREKEEPING, RETURNED AND RESUBMITTED PACKETS	.1	.2
SUPPLEMENTAL TESTS	.2	.4
COMPUTER EDIT	.05	.15
PRO-RATED COMPUTER PROGRAM DEVELOPMENT*	.05	.3
PRO-RATED MANUAL EDIT PROCEDURES DEVELOPMENT**	.10	.5
	.90	2.35

---

* BASED ON 100 HOURS OF EFFORT SPREAD OVER 2000 VEHICLES FOR PROGRAM 1 AND 120 HOURS OF EFFORT SPREAD OVER 400 VEHICLES FOR PROGRAM 2.

** BASED ON 200 HOURS OF EFFORT SPREAD OVER 2000 VEHICLES FOR PROGRAM 1 AND 200 HOURS OF EFFORT SPREAD OVER 400 VEHICLES FOR PROGRAM 2.

Table 11

CAB Test Programs (April, 1977 - April, 1978)

Name	Total No. of Vehicles tested	Total Contract Cost ($)	Vehicles Edited 4/77 - 4/78	Prorated Cost 4/77 - 4/78	In-house Edit Effort (Man-hours)
FY75 EFP	2220	$1,600,000	800	$580,000	720.0
FY77 EFP	2329	$2,000,000	1400	$1,200,000	1622.5**
6000-8500 Lb Trucks	250	$200,000	250	$200,000	587.5
FY75 Motor Cycles	250	$175,000	250	$175,000	375.0
FY75 R.M.	400	$700,000	200	$350,000	470.0
FY77 R.M.	100	$150,000	100	$150,000	235.0
Portland	6400	$2,700,000	4000	$1,810,000	4000.0
	11,949	$7,530,000	7000	$4,465,000	8010.0 (4.7 mys)*

* 1700 man-hours/year (This figure is between the 1570 hours/year DPM figure and the 1850 hours/year figure which assumes only five weeks of holidays, vacations, and sick leave.)

** Include 250 RM sequences.

Table 12

FUTURE APPROACHES - CURRENTLY BEING TESTED ON A TRIAL BIAS

1. MORE AUTOMATED DATA ACQUISITION
   - A. DEDICATED COMPUTER ON SITE (GENERATE DRIVING TRACE, SET DYNAMOMETERS, AUTOMATIC DATA RECORDING).
   - B. CENTRALIZED COMPUTER.
   - C. COST IS HIGH ($75K - 125K FOR A DEDICATED SYSTEM) AND DIFFICULT TO JUSTIFY FOR YEAR AT A TIME CONTRACTS.
2. REQUIRE STRICTER CONTRACTOR DATA EDIT PROCEDURES
   - A. ALL KEYPUNCHED DATA MUST BE ENTERED AND THEN VERIFIED BY TWO DIFFERENT PEOPLE.
   - B. REQUIRE CONTRACTOR TO APPLY MANUAL AND COMPUTER EDITING TECHNIQUES.
   - C. CHECK-OUT OF ALL CONTRACTOR DATA HANDLING PERSONNEL.
   - D. CONTRACTOR WILL SUBMIT ERROR PRINT-OUT WITH EACH GROUP OF TEST PACKETS.
3. STATION EPA PERSONNEL FULL-TIME AT EACH TEST SITE
4. ASSUMING CONTRACTOR ERROR RATE DECREASES, USE SPOT INSPECTION OF PACKETS
5. USE STRATIFIED SAMPLING SPOT INSPECTION OF PARAMETERS TO MINIMIZE COST OF ERROR TIMES VARIANCE IN VARIABLE J. STRATA CAN BE DIFFERENT DRIVERS, ETC. EACH STRATA FOR EACH VARIABLE IS INVERSELY PROPORTIONAL TO THE SQUARE ROOT OF ERROR COST AND DIRECTLY PROPORTIONED TO THE VARIANCE OF THE MEASUREMENT IN THE STRATA. THIS IS ONLY GOOD FOR CORRECTLY ESTIMATING THE MEAN OF VARIABLE J.
6. SEQUENTIAL LIKELIHOOD RATED TEST. BASED ON MINIMIZING THE COST ASSURE A GIVEN OVERALL ERROR RATE. START BY EDITING THE VARIABLE WITH THE HIGHEST IMPACT ON THE EMISSION RATE OR THE SMALLEST COST/BENEFIT RATIO IF COSTS OF EDITING ARE DIFFERENT.

Table 13

SPOT TESTING

QUESTION: GIVEN A TOTAL TEST POPULATION OF N CARS, IF WE EDIT A SAMPLE OF Y CARS AND FIND NO ERRORS, WHAT IS THE LIKELIHOOD THAT THE ERROR RATE IN THE ENTIRE POPULATION IS LESS THAN X%.

ASSUME UNKNOWN ERROR RATE OF P%, SAMPLE SIZE N, TOTAL NUMBER BAD TESTS X = PN, TOTAL NUMBER OF GOOD TESTS IS N-X = Y.

PROB. OF N GOOD PACKETS IS

$$P = \frac{Y}{N} \frac{(Y-1)}{(N-1)} \quad \frac{(Y-2)}{(N-2)} \quad \frac{(Y-N+1)}{(N-N+1)} = \frac{Y!(N-N)!}{(Y-N)!N!}$$

SEE P (CONFIDENCE LEVEL), N, P (ERROR RATE)

DETERMINE N

Table 14

Level of Certainty of Error Estimation, P

For a total population of 100 tests

Number of samples = n

Number of erroneous tests = x	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	50	60	70	80	90
1																		60	70	80	90
2															51	64	75	84	91	96	99
3														58	66	79	84	94	97	99	
4													60	69	77	88	94	98			
5												56	68	77	84	93	97	99			
6										51	55	63	75	83	89	96	99				
7									53	57	60	69	80	88	93	98					
8								54	58	62	65	74	84	91	95	99					
9							54	59	63	67	70	78	88	93	97						
10						53	58	63	67	71	74	82	90	95	98						
11					51	57	62	67	71	74	77	85	93	97	98						
12					55	60	65	70	74	77	80	87	94	98	99						
13				51	58	63	69	73	77	80	83	90	96	98							
14				54	61	66	71	76	80	83	85	91	97	99							
15				56	63	69	74	78	82	85	87	93	97								
20			60	68	75	80	84	88	90	93	94	97	99								
25		58	69	77	83	88	91	93	95	97	98	99									
30	51	66	77	84	89	93	95	97	98	98	99										

≤50% Certainty

≥99% Certainty

Interpretation:
If one finds no errors among n randomly selected test packets, one can conclude, with P% certainty, that there are no more than x erroneous packets in the total population.

For example, if 40 packets of the original 100 are examined and no errors discovered, there is only a 1% chance that there are 8 or more errors in the total group of test packets, while there is a 7% chance that there are 5 or more erroneous packets.

An entry of 99% implies that for the presumed error rate x, the chance of drawing n perfect tests is ≤1%.

DATA VALIDATION TECHNIQUES USED IN MOBILE SOURCE TESTING

by

C. Don Paulsell

Office of Mobile Source Air Pollution Control
U.S. Environmental Protection Agency
Ann Arbor, Michigan 48105

INTRODUCTION

The EPA laboratory at Ann Arbor is the primary government facility responsible for certification testing of engine-driven vehicles to determine compliance with the standards for emissions levels and fuel economy. Approximately 2500 to 3000 vehicles of foreign and domestic manufacture are tested annually. This testing is performed in 10 dynamometer test cells using the constant-volume sampling (CVS) method to collect emission samples from vehicle exhausts. The samples are analyzed on seven analyzer sites each equipped with all of the various instruments necessary for sample analysis. As the vehicles are operated through a prescribed simulated driving cycle, sufficient data are also recorded to determine fuel economy. A complete data set for a vehicle includes information such as vehicle identification data, test specifications, instrument calibrations, calibration data correlations, test data, calculated (reduced) test data, vehicle manufacturer's test results, EPA test results, and quality control data. After these data have been collected and/or generated, they are subjected to quality control procedures to assess overall accuracy, precision, uniformity, and validity.

QUALITY CONTROL SYSTEM

The "products" of our test process are the data which represent the intangible exhaust emissions of the vehicles tested. The quality control system assesses the acceptability of this product in terms of accuracy, uniformity, and validity.

Accuracy is important since these data are used to decide whether a vehicle meets federal standards. Moreover, a financial penalty may be applied to any manufacturer for not meeting the standard for fuel economy. This assessment is five dollars per vehicle produced for each tenth of a mile per gallon less than the standard. Thus, the question of accuracy could potentially involve millions of dollars.

Since accuracy can be a relative attribute, the data are also checked for precision and uniformity to determine whether measurements can be repeated at an analyzer site and whether results from each of the seven analyzer sites are essentially equivalent. Finally, since the "data product" is dependent upon the test process used, the validity of that process must be verified. Data validation techniques comprise a very important part of the total quality control system.

TYPES OF DATA VALIDATION

Data validation begins long before the vehicle test is performed and continues after the vehicle has been returned to its company.

This broad application of data validation is illustrated in the five parts of the overall process.

1. Calibration Acceptance
2. Operational Verifications
3. Procedural Checks
4. Test Data Review
5. Comparative Measures

The following paragraphs discuss each of these areas and provide examples of the methods used.

CALIBRATION ACCEPTANCE

A wide variety of instruments and equipment are used in the measurement process. It is obvious that each unit must be calibrated, but what is not obvious is how to validate that a calibration has normal characteristics. The calibration procedures are often more complicated than the test procedures - an erroneous calibration can only produce erroneous test data.

The QC methods used for calibration validation emphasize the quantitative aspects of the equipment characteristics. For example, a dynamometer is calibrated to establish residual bearing frictions. These frictional values tend to have predictable magnitudes across all dynamometers. Use of this characteristic can provide confidence in both accuracy and uniformity for dynamometer calibrations.

The analyzer and constant volume sampler also have unique characteristics. An analyzer curve can be assessed in terms of nonlinearity, curve fit deviations, and the absence of inflections. A CVS utilizes a critical flow venturi which has a characteristic discharge coefficient of .985 to .995. This coefficient, the ratio of actual to theoretical flow for a given throat diameter, can be used to assess flow metering accuracy and long term stability.

The dynamometer, constant volume sampler (CVS), and analyzer represent the three major components of the measurement process. A proper calibration is a necessary condition for getting valid test results, but the operational verification is equally important.

OPERATIONAL VERIFICATIONS

This phase of the process is used to assure that the equipment can measure and produce a known result. Special tests are conducted at daily, weekly, or bi-weekly intervals to produce a QC parameter which can be normalized relative to all systems. These parameters are manipulated statistically or plotted graphically to assess control of the process accuracy and precision.

For example, the CVS is checked by injecting a known mass of pure propane as though it were auto exhaust. All measurements and calculations are performed as in a test and the result must be within plus or minus 2 percent of the known value. Leaks, calibration drift, erroneous analyzer span gas values, and many other parameters can cause the verification to fail.

The analyzer is verified daily by analyzing a bag of blended gases at each of the seven analyzer sites. The deviation of each site from the overall average serves as the normalized parameter. A site which is consistently high or low or inconsistent will be obvious from the automated control chart analysis. Positive or negative consecutive runs greater than five, or excessive data scatter are automatically flagged and noted by a QC message on the analysis printout.

The dynamometer gets a short version of the full calibration to verify its stability. Control charts of flywheel frictional values will graphically show a deteriorating bearing or load control problem.

Finally, a repeatable car is tested on each site utilizing all the normal components of the system. The emission results are statistically analyzed to show significant differences between sites. These operational verifications each address a specific part of the process, and when assessed in total, provide assurance that the system is capable of producing valid emission results from a properly conducted test procedure.

PROCEDURAL CHECKS

A complete emissions test can require a total of about eighteen hours, including the twelve hour overnight "soak" period. The specifications and criteria are so numerous that a set of checklists have been developed to document that each one has been done properly. Test times, temperatures, shift patterns, horsepowers, special procedures, and many other conditions are noted or checked off as each phase progresses. In some cases, such as fueling, the operation must be witnessed by two people, since the type of fuel can greatly affect emissions.

Although the test equipment has been previously verified, several checks are performed as part of the test. An open valve or improper horsepower setting would cause the test to be voided.

At the end of the test process all the stripcharts, checksheets, datasheets, and driving traces are consolidated, reviewed, and sent for computer processing.

TEST DATA REVIEW

The test processing office validates that all necessary data have been obtained. The data sheet is then batch processed by computer to generate a printout of input data, calculated results, QC checks, and pass/fail criteria. The computer program has been designed to audit the various test data for omissions or unrepresentative values.

For example, a 40,000 pound automobile would likely be a 4,000 pound value improperly keypunched. Other data, such as ambient background concentrations can be compared to a normal distribution of values obtained at EPA to flag high levels. Higher values may indicate improper analyzer parameters or a leaking vehicle exhaust system. Since some of the test sequence is repeated, a ratio of two flowrates or distances travelled can be very useful in highlighting abnormalities. A normalized ratio has become a valuable tool because it is not affected by the magnitudes of parameters, which may normally be different. It is the ratio of these different magnitudes that produces a value which lies within a narrow bandwidth. The ratio of highway to city fuel economy is an example of this application.

If all data have been validated and all acceptance criteria met, the documentation is stored and the results are updated as valid in the computer data base. While this completes the processing of one test, it is not the end of the data validation process.

COMPARATIVE TESTS

Each test alone has certain characteristics, and all tests combined have other useful measures. Comparative tests on large populations of vehicle results can highlight differences and trends that an individual test does not show.

The manufacturer has normally tested the vehicle prior to EPA's test, so an independent set of data is available for comparison. The MFR/EPA emission differences and percent differences are calculated and stored in a "paired data" file. These normalized values can then be statistically summarized for each manufacturer group. The results of this analysis show the relative agreement between EPA and all individual manufacturers. If EPA is consistently higher or lower, a systematic bias may be indicated. Diagnostic tests or correlation programs can be performed to identify and correct the cause.

Statistical analysis of all these data can provide the upper and lower limits which are used to assess the significance of a bias. Test conditions and equipment identifiers can be used to stratify the analysis for assessing whether such things as altitude differences or specific test sites correlate with the paired data differences. Finally, the data validation loop can be refined by the statistical determination of QC limits.

QUALITY CONTROL REFINEMENTS

A strong data validation program can be developed by automating many of the checks being made. Computerized validation and acceptance tests require that the data be pertinent and accessible. Use of an integrated data base structure can minimize manual operations, improve security, and assure the integrity of the data.

A computerized data base can also enable the automation of screening programs, plotting routines, and statistical summaries. It will permit rapid development of more precise tools and tests which can be used in the data validation process.

Finally, a computer data base can provide a trail for audits or requests for documentation.

CLOSURE

This paper has shown that the data validation process is not simply an inspection of results at the end of a test. Rather, it is a combination of specific individual tests and checks which when taken as a whole, form the foundation for a quality control system which can provide documented, quantitative assurance that the "data product" of the EPA mobile source program is fit for use in our regulatory process.

VALIDATION OF CONTINUOUS STACK MONITORING DATA

by

Joseph E. McCarley

Emission Standards and Engineering Division
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

VALIDATION OF CONTINUOUS STACK MONITORING DATA

J.E. McCarley

SUMMARY

The Emission Standards and Engineering Division is currently developing a revised standard of performance for new steam generators. As part of this study, the feasibility of continuous regulation of sulfur dioxide emissions, as well as a percentage of sulfur reduction from fossil fuels, is being evaluated. In support of this study, the Emission Measurement Branch is conducting sulfur dioxide continuous monitoring projects at five coal-fired power plants equipped with flue-gas desulfurization units. When data are being collected for supporting regulations, validation of these data is an important consideration.

Prior to collecting emission data, the continuous monitoring systems are validated by following the procedures described in Performance Specifications--Appendix B 40 CFR 60. (Performance Specification 2--Performance Specifications and Specification Test Procedures for Monitors of SO_2 and NO_x from Stationary Sources.)

The monitoring data are then collected and recorded continuously from each emission point at least once every 15 minutes. In this study, the data are then placed in a computer bank, printed and then edited or validated manually. During the monitoring periods when data are collected during instrument malfunction, calibration, or plant upset conditions the time periods for these conditions are recorded by plant personnel. These data are purged from the computer bank and the remaining data are averaged for each 1-hour, 3-hour, 8-hour, 24-hour, and 30-day periods of time. If more than one 15-minute data point has been determined to be invalid in any one hour period, that entire 1-hour data are considered invalid and not included in the longer averaging periods. In summary, the data are edited for actual known errors and no

statistical validation procedures are performed.

Further details of these monitoring projects are contained in the following report and references therein: Kelly, W. and Sedman, C. First Interim Report: Continuous Sulfur Dioxide Monitoring at Steam Generators. EMB Project No. 77SPP23A, Emission Standards and Engineering Division, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, June 1978. 54pp.

Future plans for evaluating validation procedures include (1) application of more automatic recording and data validation instrumentation, and (2) quality control steps to assure the accuracy of long-term emission monitoring.

SCREENING CHECKS USED BY THE NATIONAL CLIMATIC CENTER

by

William E. Klint

National Oceanic and Atmospheric Administration
National Climatic Center
Federal Building
Asheville, North Carolina 28801

SCREENING CHECKS USED BY THE NATIONAL CLIMATIC CENTER

W.E. KLINT

ABSTRACT

Current processing is discussed with emphasis on validation checks and manual interface. The need for an automated quality control program is recognized and plans for such are presented. Plans for a new modular surface edit are presented along with a new quality control procedure using an interactive graphics system. Data management is addressed through a Data Dictionary/Data Base Management system.

The National Climatic Center is:

> Responsible for receipt, processing, archiving and publication of climatological data. Coordinates the analysis of past meteorological data for NOAA, other Government agencies and the public to accommodate user requirements for climatological data through special studies and statistical analyses. Manages the national program of climatological data recall and works closely with the military in meeting this special requirement. Provides facilities, data processing support, and expertise, as requested, for World Meteorological Organization programs (e.g., GARP and GATE).[1] Assists in training programs to familiarize the representatives of developing countries with modern meteorology and coordinates (through World Data Center-A) international exchange of climatic data.

Of the various types of incoming data, paper forms predominate. These then must be keyed to digital form for processing. This effort entails keying approximately 37 million bytes of data per month. Because of a cutback in funding years ago only three-hourly surface observations, or eight observations per day, are digitized.

At the present time, processing exists in two modes; a machine edit, and a manual interface.

The machine edit consists of data verification, a range limit check, a cross-field consistency check, a continuity check, and appropriate flags to "verifiers."

Data verification is a simple machine check to see if there is indeed data keyed into the appropriate field, and if there, and the field is coded, is it a legitimate code.

The range limit checks to see if the value in a particular field falls into an appropriate range. However, at the present time there is only one range limit per field. This, in and of itself, causes many unnecessary "kickouts."

The cross-field consistency check looks at the entries in related fields for consistency; i.e., clouds and precipitation.

The continuity check does a range limit check on certain fields between the previous observation and the one being checked.

Finally, if any of the above checks fails, the appropriate flags are printed out for return to the "verifiers" and appropriate action.

The manual interface, due to the magnitude of data, consists primarily of a visual scan of all forms. A random sampling of stations receives a closely scrutinized check of all observations. Problems with the data requiring corrections are handled as follows: first, the erroneous entry is crossed through with a blue pencil and the "correct" entry is made directly above the erroneous one. Second, if the observation is one which is normally digitized, a change form is routed to key entry.

The "kickouts" from the machine edit, which were returned for action, are scrutinized, a decision on validity is made, and, if necessary, a correction is made both on the original paper form and on a change form to key entry.

The change forms are routed to key entry for digitizing and the changes are again run through the machine edit.

The above procedure is a recurring one until no more errors appear. Once all the data "pass" the edit, they are formatted into the surface observation file and entered into the data bank.[2]

It is fairly obvious that, due to the rather limited nature of these checks, some erroneous data slip through and are placed into the data bank. This fact, coupled with the realization that the magnitude of incoming data in digital form is on the increase, and with the fact that a more closely "real time" edit is both possible and needed, is forcing changes upon NCC.

Although the basic processing stages of machine edit and manual interface will remain the same, the nature of each will take on a new and challenging meaning.

With the innovation of the new National Weather Service Automation of Field Operations and Services (AFOS) system, the NCC will acquire near real time collection capabilities of data in digital form. These, plus manuscript forms, create a real need for dual processing of data.

The edit computer program is being completely rewritten, as in its present form it is difficult to maintain. It is designed in a modular form and many previously manual functions are designed into the program.

The creation of a Master Station Inventory (MSI) will completely change the complexion of the edit program. The basic edit routines remain the same, with the following changes:

1. The verification step will now be checked against the MSI for validity. Previously some missing entries were flagged to a "verifier" whether they were missing or simply not observed at that particular station. The MSI will now be checked for proper disposition before an error flag is returned, thus alleviating the "verifier" of this task.

2. The creation of the MSI will allow for a complete set of range limits for every field of every individual station, thus preventing unnecessary "kickouts" for "good" data, and providing for a narrower range limit check of each field.

3. Cross-field consistency checks will remain basically the same with the provision that with the above mentioned checks, should be more reliable. They have been "beefed-up" to contain closer checks and checks previously left to the "verifiers."

4. If an error is isolated and a flag is called for, a check is first made with the MSI to see if a mathematical relationship exists. If one does, a new value is calculated and entered beside the original with an appropriate flag.

If an error is isolated and no mathematical relationship exists, the appropriate flag is issued and the observation queued for scrutiny by a "verifier." All observations changed by a "verifier" are automatically re-entered into the edit program.

The manual interface by the verifier will consist of interacting with the data through use of an interactive graphics system. The "verifier" previously had only manuscript forms as input to his decision. Now he will be able to present the data in any of several displays including contoured map analyses of a surrounding areal coverage. With this input the verifier will be able to make a more intelligent decision as to proper disposition of questionable data.

Up to this point we have discussed only a superficial edit of the incoming data. We have not, as yet, looked at the inherent quality of the data itself. NCC, at the present time, does not have the capability of doing relational checks on the data. With the acquisition of the Asymptotic Singular Decomposition (ASD) model, developed by Dr. John Jalickee,[3] CEDDA, NCC now has this capability.

In its simplest terms the ASD model uses the method of least squares on a data matrix.

The first step is to calculate a "characteristic" vector for the matrix. Next, the differences between the data matrix and the appropriate "characteristic" are calculated. The matrix is now overlayed with these differences and the process is iterated.

The first component of vector magnitudes, when plotted, results in a graph of the dominant features; the second component, the features of the difference matrix, etc. We have found that with most data fields, the second and third component plots prove to be the most useful for validation. By the time the fifth component plot is made, we have usually reached the noise level.

The data, thus plotted, can be expected to show "continuity." The physical relationship of the field should be apparent in the graph. If that relationship breaks down at any point in the graph, we can assume bad data. This model will give NCC the capability to perform qualitative (relational) checks on all incoming data.

A side effect benefit of this model is the capability of building a station "normal" situation. Based on this, such things as instrument drift, miscalibration, and erroneous launch data become readily apparent when exposed to a trend analysis. Once isolated, these "bad" data can be adjusted "toward" the normal with at least some degree of accuracy.

The concept of the "verifier's" job changes somewhat under this new approach. The computer edit now will do much of the job the verifier did previously, thus relieving him of that task; the bulk of which was scanning "good" data. Upon his arrival for duty, he sits down at a KCRT console or terminal and calls up the flag file for his particular area of interest. He sits in the seat of judgment and makes those decisions too delicate or volatile to have been programmed into the edit routine. Once made, these observations are returned to the edit queue to be run once again. Only after an observation "passes" the edit program is it allowed to continue into the ASD model.

The results of the ASD run are displayed in one of several graphic modes for verification. Realizing that the normal range of this display is from +0.5 to -0.5 units, its power and usefulness becomes apparent. Remember here that this is a display of the second or third component of the data field, and, as we are working with differences, should nicely fit within this range. The "outliers" will stand out here with striking notoriety. The verifier now has the task of "replacing"* the "outliers" with a more reasonable value. This can be done simply by sight alignment of his cursor or light pen with the trend of the curve or by having the computer do a best fit. Although this sight alignment appears to be a rather gross correction, when it is "blown up" into the initial state it becomes very tolerable.

All original data are kept, with corrections and appropriate flags being entered adjacent to them before being incorporated into the NCC data bank. This will allow use of either datum by the user.

The NCC is currently planning a database environment. This quality control process will allow us to place only QC'd data of a high reliability into our database, thus assuring the user of quality data. Another side effect of ASD is its compaction possibilities for storage. The set of components for a data field can be "blown up" to explain 99% of the original field; thus NCC can store components and blow them up to the "original" field on output. This will result in many orders of magnitude reduction of the necessary storage facilities.

---

*Note here that "replace" does not imply that we destroy the original value. It will be maintained and output along with the corrected value.

REFERENCES

For your convenience, a copy of the following three references are included herein, starting on the next page. The generosity of Mr. Walter James Koss, Primary Data Branch, EDS, Asheville, NC 28801, for supplying these references for publication in these Proceedings is appreciated.

1. Barton, G. and Saxton, D. The Role of Interactive Computer Systems in Data Processing at CEDDA. Environmental Data Service (EDS) Magazine, pp. 10-14.

2. Edit Procedures - Surface Observational Data. Surface Section, Primary Data Branch, National Climatic Center, Asheville, NC 28801. August 1975, 31 pp.

3. Jalickee, J., et.al. Validation, Compaction, and Analysis of Large Environmental Data Sets. Environmental Data Service (EDS) Magazine, pp. 3-9.

The Role of Interactive Computer Systems in Data Processing at CEDDA

By Gerald Barton and David Saxton

Introduction

The Environmental Data Service's Center for Experiment Design and Data Analysis (CEDDA) processes enormous volumes of interdisciplinary environmental data collected in major field research programs and projects, such as the recent GARP (Global Atmospheric Research Program) Atlantic Tropical Experiment (GATE). As an example, CEDDA received 1,700 miles of magnetic tape data from the four U.S. ships (*Researcher*, *Oceanographer*, *Dallas*, and *Gillis*) in GATE's primary array.

CEDDA's goal is rapid processing to provide the data to the scientific community as soon as possible after the completion of a field experiment. One necessary step is editing the data to remove invalid readings. CEDDA's current turnaround time for interactive editing of a data file is 1 to 3 weeks. It is hoped that a new interactive computer system CEDDA is currently assembling will cut this time to ½ hour or less.

Data Collection

During field experiments, environmental data are recorded continuously by instruments on ships, towers, buoys, balloons, and other platforms at sample rates from 10/second to 4/second. A wide variety of specially calibrated sensors measure such variables as temperature, dewpoint, pressure, wind, radiation, salinity, and rainfall. The outputs are processed and stored on multitrack magnetic tapes. One track is used exclusively for time so that the exact Julian date, hour, minute, second, and 1/10 second for each sample are known.

To augment this high-resolution taped data, each major sensor subsystem output is supplemented by logs, stripcharts, and optical marked cards that record calibration checks, sensor changes (with all serial numbers), and special events, such as the beginning or end of an instrument cast.

The completeness of the data sets and their security are matters of prime concern. At the end of a phase of a field experiment, or at other convenient intervals, all tapes, logs, cards, etc., are shipped to CEDDA using the safest methods available. During GATE, CEDDA had a data manager on each of the 4 U.S. ships in the primary (B-scale) array and also at the GATE Operations Control Center to ensure the completeness and security of the transfer process.

Current Processing System

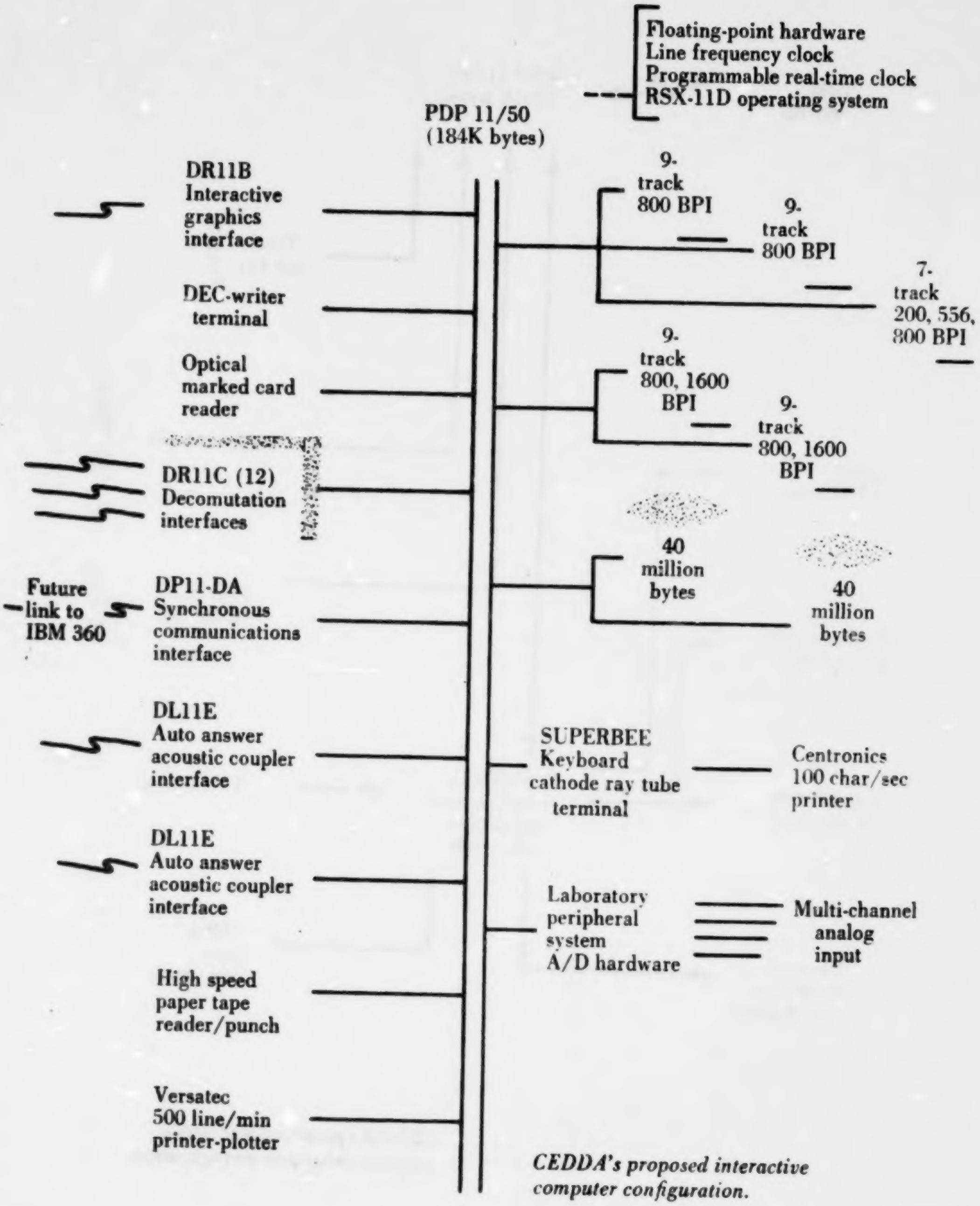
At CEDDA, the incoming analog data tapes are first checked for recording quality and completeness. Next, a minicomputer converts the analog data to digital form, producing a digital tape. Playback time is 32 times faster than field recording speed, so an 8-hour field tape is transcribed in about 15 minutes. During the minicomputer processing, an additional computer time word is added to each sample to control subsequent data processing programs and to preclude the loss of any sensor data due to malfunction or noise in the field time system.

Processing next proceeds to one of NOAA's larger computer systems, where data sets are organized by component systems used on the data collection platform, e.g., Oceanographic Data Set or Rawinsonde Data Set. Graphical display of the data as time-series plots and graphs, and frequency distribution plots, is required for the analysis of these data sets.

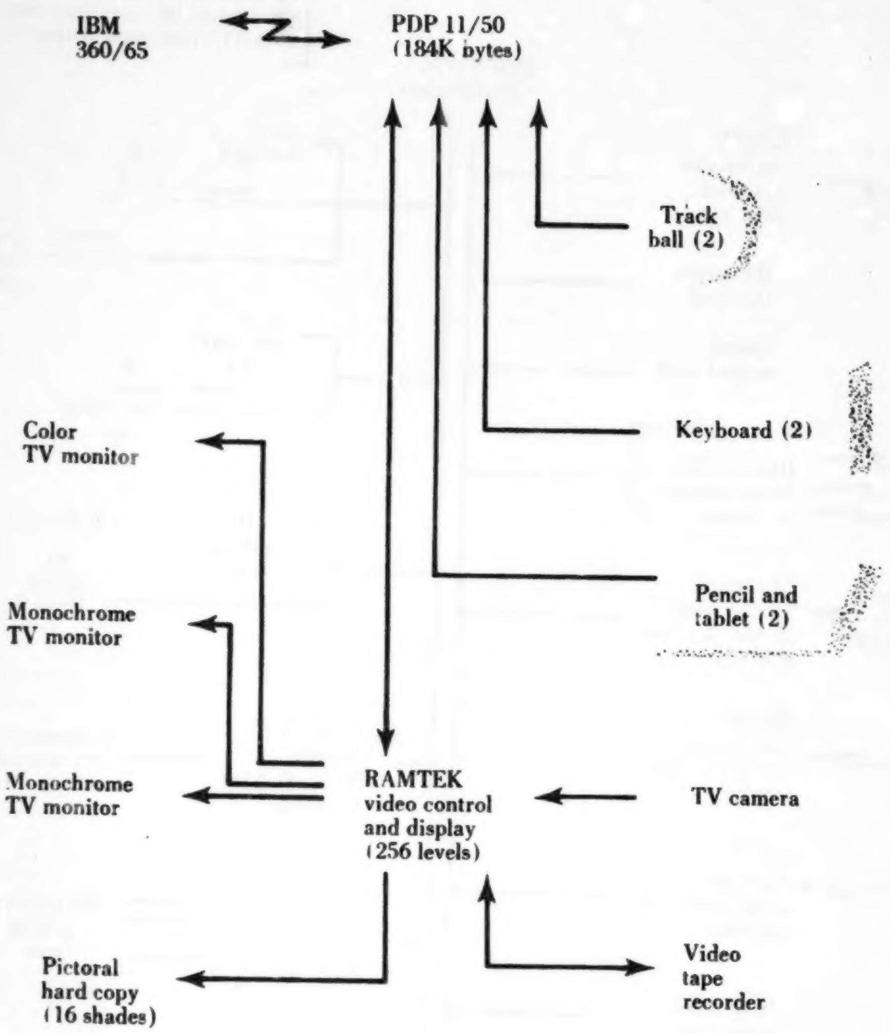
The editing features of the current computer processing system can be thought of as an interactive graphics system, with the time required for interaction varying up to a week or more. For optical mark cards, reaction is rapid since all event cards may be listed in chronological order and cards may be inserted, deleted, or corrected using a list-edit program in the minicomputer. However, for high-resolution meteorological or oceanographic data which must be transformed to engineering units and properly scaled, display for editorial review is currently limited to a microfilm graphics subsystem located in nearby Suitland, Md. For these data sets the time required for interaction includes the transport of data tapes, generation of microfilm graphics in a batch mode at the remote site, transport of microfilm on the return loop, review using microfilm readers, testing of automated corrections when required, and the recycling to display

New Processing System

CEDDA is currently assembling the hardware and software necessary to implement an interactive computer system that will allow the data editing and updating functions to be performed in a single processing step (real time). The main components of the system will remain a Digital Equipment Corporation (DEC) PDP-11/50 minicomputer and an IBM 360/65. It will be possible to access data on the IBM 360/65 through the PDP-11 or through terminals. The


Floating-point hardware
Line frequency clock
Programmable real-time clock
RSX-11D operating system
PDP 11/50 (184K bytes)
DR11B Interactive graphics interface
DEC-writer terminal
Optical marked card reader
DR11C (12) Decomutation interfaces
Future link to IBM 360
DP11-DA Synchronous communications interface
DL11E Auto answer acoustic coupler interface
DL11E Auto answer acoustic coupler interface
High speed paper tape reader/punch
Versatec 500 line/min printer-plotter
9-track 800 BPI
9-track 800 BPI
7-track 200, 556, 800 BPI
9-track 800, 1600 BPI
9-track 800, 1600 BPI
40 million bytes
40 million bytes
SUPERBEE Keyboard cathode ray tube terminal
Centronics 100 char/sec printer
Laboratory peripheral system A/D hardware
Multi-channel analog input

CEDDA's proposed interactive computer configuration.


IBM
360/65
PDP 11/50
(184K bytes)
Track ball (2)
Keyboard (2)
Color TV monitor
Pencil and tablet (2)
Monochrome TV monitor
Monochrome TV monitor
RAMTEK video control and display (256 levels)
TV camera
Pictoral hard copy (16 shades)
Video tape recorder

CEDDA's proposed interactive graphics subsystem configuration.

Reproduced from
best available copy.

PDP-11 will have a graphics subsystem that will take less than 30 minutes to perform the functions of the current microfilm subprogram.

The major features and components of the interaction system are:

(1) Access to the IBM 360/65 timesharing facilities via keyboard cathode ray tube (KCRT) terminal, ASR-33 teletype terminal, or PDP-11/50 minicomputer.

(2) Input terminals to the PDP-11/50, including an LA-30 DEC writer terminal, a KCRT, and two dial-in terminal interfaces for use with remote terminals.

(3) A graphics subsystem for the PDP-11/50.

(4) DEC's (RSX)-11D real time, priority-driven, multidisking executive system for the PDP-11/50.

With these features, a user can access the 360 to perform mathematical computations or generate data sets. He can look at the data and analyze them in real time on the interactive graphics subsystem. When he finds errors, he can immediately correct the data, and display them again on the graphics system to validate the corrections. He can then archive the updated data set for future use.

Interactive Graphics Capabilities

The interactive graphics subsystem, designed and assembled by Operating Systems Incorporated of Tarzana, California, consists of a RAMTEK graphics display system interfaced with CEDDA's PDP-11/50 computer by an appropriate switching network. Features of the full system (only part of which is required for the data editing job) include two black and white TV monitors, one color TV monitor, two data entry keyboards, two pencil and tablet systems, two track ball cursor controls, a television tape recorder with microphone input, a TV camera with zoom lens, an analog to digital convertor, eight planes of memory that allow up to 256 shades of gray or color and a cross-print switching network that allows mixing control of inputs and outputs.

A simple use of an interactive graphics system is the editing of raw data displayed as a time-series analysis or plot. For example, a single parameter, such as temperature, is plotted at its highest resolution in a time sequence covering several hours or days. Visual inspection of the data may reveal large errors where the sensor or telemetry system failed. To correct these larger errors, a window edit program might be tested with all "good" values of the parameter constrained to fit between the upper and lower limits of the window. Diurnal and other trends might be superimposed on the data plot. The limits and trends can be displayed with the raw data to show which data points should be edited out.

A slightly more sophisticated version of this time-series plot would compute running means over minutes or hours and show which of the high-resolution points will fall outside two or three standard deviations. Complex curves using higher order polynominals can be fitted to time-series data, both before and after various editing passes, to eliminate, insofar as possible, "noise" from the data. Various filters and smoothing functions also can be tested and evaluated before going into an Automatic Data Processing (ADP) production mode.

In general, CEDDA's new interactive graphics system will make it possible to display two or more curves simultaneously, using color, intensity, or blinking characteristics to distinguish, for example, between a standard and trial edit scheme or between different parameters. It will provide the capability to produce hard-copy documentation of both the trial programs as they progress during a test and the data sets used.

A more demanding requirement of an interactive graphics system is the ability to display and operate on digitized field data. An example of this type of data is a digitized radar picture. Under the control of an interactive graphics system, the analyst should be able to select and display a radar picture, to rotate and rescale it to a standard grid size, to enhance the digitized increments by contours or false color transfers, to overlay and compare it with the previous picture, and to display only those points from the two pictures whose change exceeds some threshold value. Similarly, the analyst should be able to display the overlap portion of digitized radar pictures from two locations and to scale and normalize these independently so that compatibility is established on common echo systems.

A further refinement is the addition of a TV-type scanner so that analog material can be rapidly digitized at high resolution and then handled with all the capabilities of the interactive graphics system. For example, a satellite visual range photograph could be scanned and digitized and then displayed with a radar picture covering the same area. Specific rainfall rates from surface observations could be overlayed on the same display so that some integration of areal rainfall amounts would be immediately available.

An interactive graphics system provides the ability to overlay data from different platforms or different systems. For example, the temperature and vertical velocity from sensors at several levels on a tethered balloon system could be compared by an analyst for coherence and lags as convective plumes are sampled. Dropsondes (atmospheric soundings) from aircraft could be graphically superimposed on simultaneous radiosonde soundings from ships. Spectra taken by instrumented aircraft during ship flybys can be compared with high-resolution data recorded on board each ship.

CEDDA plans to have the new interactive computer system in operation by late 1975. By that time, implementation of the graphics subsystem should include the work done in the current COM cycle. Future CEDDA applications of the graphics system will include programs that allow display of radar or satellite pictures in multicolors or up to 256 shades of gray using the tape-recording features of the graphics system. It should be possible to construct time-motion pictures of changing weather features. Also envisioned is the capability to display slices through 3-D models of weather systems. CEDDA currently has analysis programs that allow an analyst to change parameters in a weather model. The real time operation of the graphical display should allow the scientists to experiment with parameters that he may never have had the opportunity to look at previously.

It can be seen from the above examples that an interactive graphics system has broad applicability, extending from program design and test through all stages of data reduction and processing to scientific data analysis. In addition, interactive graphics provides programmers and analysts with the ability to see the data move through programs from recorded voltages on multiple channel tapes until they become validated meteorological or oceanographic data suitable for permanent archival and dissemination to the user community.

Gerry Barton

Dave Saxton

The Authors

GERALD BARTON, Chief of CEDDA's Computer Systems Branch, has a B.S. degree in Geophysics from Pennsylvania State University and an M.A. in Geological Science from the University of Texas. Before coming to CEDDA, he worked for ten years with the U.S. Naval Oceanographic Office as a geophysicist. His early association with the Oceanographic Office included gravity survey cruises in the USS Archerfish, a research submarine, in the Western Pacific and off the east and west coasts of the United States. From 1967 through January of 1974, when he joined CEDDA, Gerry spent most of his time working in computer programming, systems design, and the processing of gravity and geodetic data—to determine, among other things, the deflection of the vertical, or "which way is up."

DAVID SAXTON joined CEDDA as Chief of the Operations Division in April 1974, following a 30-year career in the Air Weather Service which took him to England, France, Germany, and Japan. Dave has a B.S. degree from the University of Michigan and an M.S. from the University of Chicago. During World War II, he served as an Air Force weather forecaster in Europe. After the war and a year of civilian/student life, he was recalled to active duty and assigned to the joint Weather Bureau/Army/Navy Weather Central in Washington, D.C. Subsequently, he was posted to the Tokyo Weather Central, then to the USAF Weather Central in Suitland, Md., later moving with that organization to Offutt AFB, Nebraska. In 1961 he was assigned as Chief of the Strategic Air Command Weather Support Center in High Wycombe, England. Four years later he was assigned to Air Weather Service Hqs., Scott AFB, Illinois, as Chief of AWS' Computer Techniques Division. In 1967, Dave returned to Offutt, now the Air Force's Global Weather Central, as Chief, Development Division, and later Chief of Operations. In 1971 he went to Hickman AFB, Hawaii, as Chief of Operations Division, Headquarters, First Weather Wing. Retiring from the military in March 1974 (with the Legion of Merit), he joined CEDDA the following month.

EDIT PROCEDURES
SURFACE OBSERVATIONAL DATA

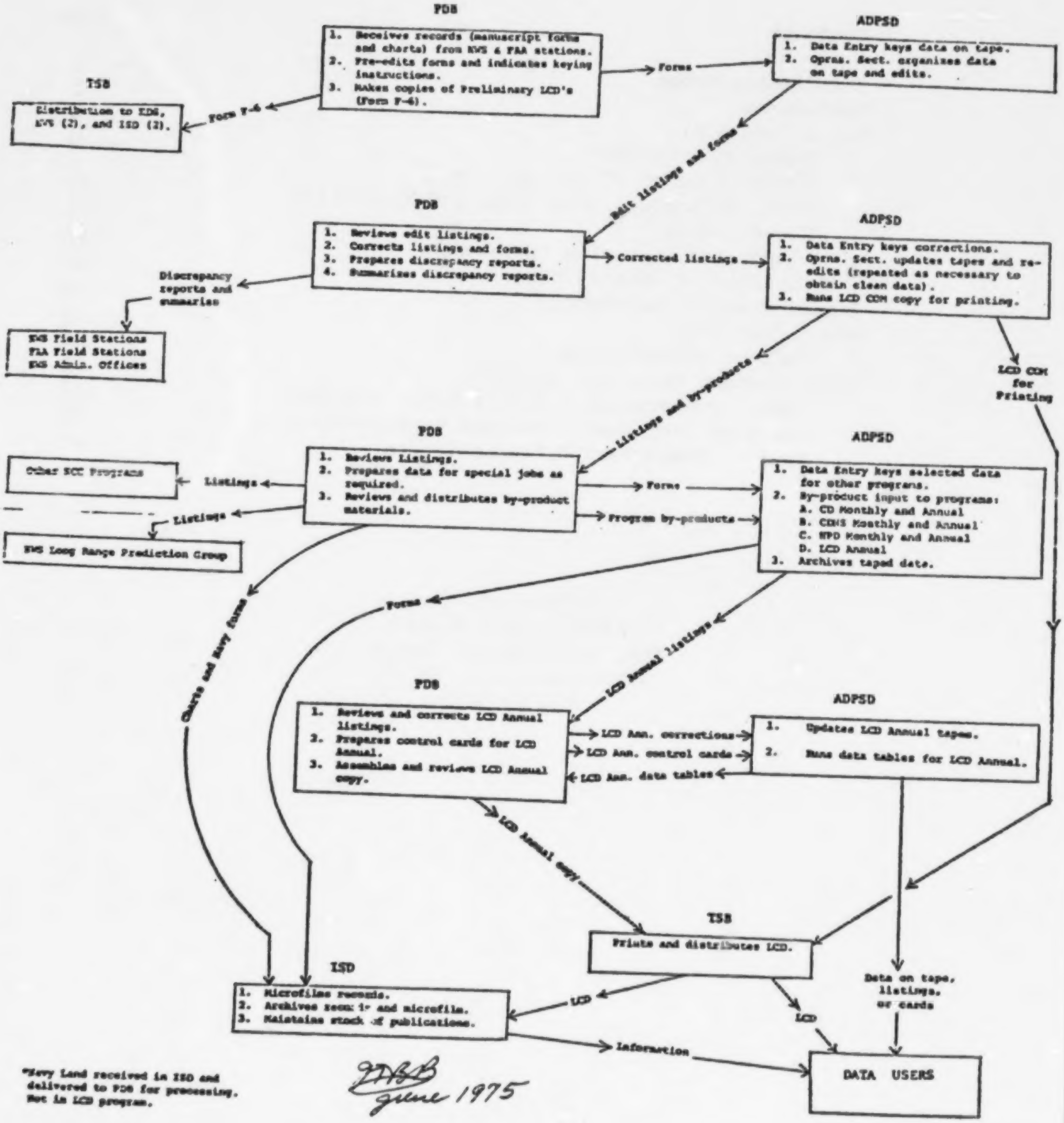
Contents

	Page
Card Images Keyed	1
Procedures	2
No. 1 Card Edit	4
Psychrometric Check	4
Limiting Range of Variability	6
Wind, Weather, Temperature, and Visibility	7
Cloud Coding	9
Clouds and Obscuring Phenomena	11
Explanation of Edit Flags	16
Visual Checking of Records	19
No. 3 Card Edit	19
Machine Computations	24
Precipitation Data Card Images	26
Checking Procedure - Hourly Precipitation	27
Checking Procedure - Extreme Precipitation	28
Maximum Short Period Precipitation	29

Surface Section
Primary Data Branch
National Climatic Center
Asheville, N. C. 28801
August 1975

SURFACE OBSERVATION RECORDS PROCESSING

NWS FAA NAVY LAND*


PDB
1. Receives records (manuscript forms and charts) from NWS & FAA stations.
2. Pre-edits forms and indicates keying instructions.
3. Makes copies of Preliminary LCD's (Form F-6).
Forms
ADPSD
1. Data Entry keys data on tape.
2. Oprns. Sect. organizes data on tape and edits.
Form F-6
TSB
Distribution to EDS, NWS (2), and ISD (2).
Edit listings and forms
PDB
1. Reviews edit listings.
2. Corrects listings and forms.
3. Prepares discrepancy reports.
4. Summarizes discrepancy reports.
Corrected listings
ADPSD
1. Data Entry keys corrections.
2. Oprns. Sect. updates tapes and re-edits (repeated as necessary to obtain clean data).
3. Runs LCD COM copy for printing.
Discrepancy reports and summaries
NWS Field Stations
FAA Field Stations
NWS Admin. Offices
Listings and by-products
LCD COM for Printing
PDB
1. Reviews Listings.
2. Prepares data for special jobs as required.
3. Reviews and distributes by-product materials.
Listings
Other SCC Programs
Listings
NWS Long Range Prediction Group
Forms
Program by-products
ADPSD
1. Data Entry keys selected data for other programs.
2. By-product input to programs:
A. CD Monthly and Annual
B. CDHS Monthly and Annual
C. HPD Monthly and Annual
D. LCD Annual
3. Archives taped data.
Charts and Navy forms
Forms
LCD Annual listings
PDB
1. Reviews and corrects LCD Annual listings.
2. Prepares control cards for LCD Annual.
3. Assembles and reviews LCD Annual copy.
LCD Ann. corrections
LCD Ann. control cards
LCD Ann. data tables
ADPSD
1. Updates LCD Annual tapes.
2. Runs data tables for LCD Annual.
LCD Annual copy
TSB
Prints and distributes LCD.
ISD
1. Microfilms records.
2. Archives records and microfilm.
3. Maintains stock of publications.
LCD
LCD
Data on tape, listings, or cards
Information
DATA USERS

*Navy Land received in ISD and delivered to PDB for processing. Not in LCD program.

June 1975

EDIT PROCEDURES
SURFACE OBSERVATIONAL DATA

I. Introduction

A. Surface records are received at NCC for processing and quality control to produce several routine summaries by machine methods from taped data. Processing includes keying, verification, and quality control procedures. After processing, records and summary products are archived at the NCC.

B. A joint machine edit program for a portion of the hourly observations has been made by EDS and AWS. However, where different, only that which is applicable to EDS is listed in this outline.

C. Data are keyed on magnetic tape. If wet bulb temperature and relative humidity values are not in the basic data as keyed, machine computations of these values are entered on tape.

D. The taped data are machine edited, corrected, and used in a number of machine programs producing various monthly and annual summaries.

II. Card Images Keyed

A. WBAN No.1 card - Hourly Surface Observations. This image is keyed only for the hours corresponding to 3- and 6-hourly synoptic times in LST for NWS, FAA, and Navy stations.

B. WBAN No. 3 card - Summary of Day. This image is keyed from the summary blocks of Form MF1-10B, the B-16 or, in a few cases, from the F-6. For FAA stations, the form is MF1-10C. In general, this image is not used when the station program is such that a summary approximately midnight to midnight is not possible.

C. The precipitation card series is:

1. Hourly precipitation - 2 images (1 & 2 keyed in col. 12) for each day of the month having precipitation and for the last day of the month with or without precipitation.

2. Maximum short period precipitation, per month - 2 images (1 & 2 keyed in col. 10) for each station per month showing maximum amounts for time intervals of 5 to 180 minutes.

3. Maximum 24-hour amounts, per month - 1 image (4 in col. 12) is keyed showing the greatest precipitation and date(s), greatest snowfall and date(s), and the maximum snow depth and date(s).

   (a) When the value is zero (0), date is left blank.

III. Procedures

A. A scan edit of the forms is made and keying instructions applicable to the station program indicated on the station folder.

B. Data Entry Section keys data on tape.

C. Operations Section transfers keyed data to computer tapes by record type.

WBAN No. 1 images, hourly observations, are placed on two tapes.

1. Tape No. 1 includes NWS (except Antarctica) and FAA stations.

2. Tape No. 2 includes NWS Antarctica and Navy stations.

3. The edit program provides for priority editing on tape No. 1 into two groups.

   a. Group No. 1, stations in the LCD program, is edited in two lots - first and second cutoffs. The first cutoff is made at the discretion of the Chief, Surface Section, when 75 to 90 percent of the records for the month are available; remaining records constitute the second cutoff. Records received unduly late can be held for processing with data for the next month.
   b. Group No. 2, stations not in the LCD program, is usually processed after completion of group No. 1.

D. WBAN No. 1 records are edited according to the station's observational program using a reference tape containing the station WBAN Number, Name, Elevation, Psychrometric Pressure Table, and Observational Pattern.

1. The observational pattern is designated by assignment of numeric values to fields in the card image and use of the sum of the field values applicable to the station for each hour as a control of the machine tests to be made.

Field	Value	Card Image Columns
Ceiling	1	14-16
Sky Condition	2	17-20
Visibility	4	21-23
Wea. & Obstruction	8	24-31
S. L. Pressure	16	32-35
Dry Bulb Temp.	32	47-49
Dew Point Temp.	64	36-38
Wind Dir. & Speed	128	39-42
Station Pressure	256	43-46
Wet Bulb Temp.	512	50-52
Relative Humidity	1024	53-55

Field	Value	Card Image Columns
Total Sky Cover	2048	56
Cloud Layers	4096	57-58
Total Opaque	8192	79

The observational pattern is keyed in two cards as illustrated in Fig. 1. The station WBAN number in cols. 1-5, the first 12 hours LST of the day in cols. 7-11, 13-17, etc., in the card keyed 2 in sol. 80 as an identifier and the last 12 hours in the card keyed 3 in col. 80 as an identifier.

2. The observations are sorted from the original tape into chronological day and hour order, edited, and one observation only for each hour (first on the original tape if multiple entries) transferred to another tape (called the sorted tape).

3. Only the records questioned in the edit are listed. Complete data, keyed and computed, in a questioned record are listed on format paper (Fig.3a) with triple spacing. Appropriate flags appear on the line above the data in the first column of the field(s) questioned. Field corrections are entered on the second line above the data for keying.
   a. An asterisk "*" indicates inconsistency.
   b. An ampersand "&" indicates data not in the station's program, except that, if there is an inconsistency, the "*" flag instead of the "&" will appear. Observations not in the station's program are edited as though all fields were required.
   c. "DUP1," "DUP2." etc., are listed to indicate duplicates up to three. All duplicates are edited, but only the first observation on the original tape is transferred to the sorted tape.
   d. "MSG" above the day and hour indicates an observation in the edit pattern is missing. "- -" for the hour indicates the entire day is missing.

4. An inventory listing (Fig. 3g) at the end of the edit listing for each station shows all hours for which observations are on tape with the total number of observations on the tape for the month at the end of the inventory listing.
   a. "01" printed under an hour indicates an observation with the cloud field keyed.
   b. "02" indicates an observation without the cloud field keyed.

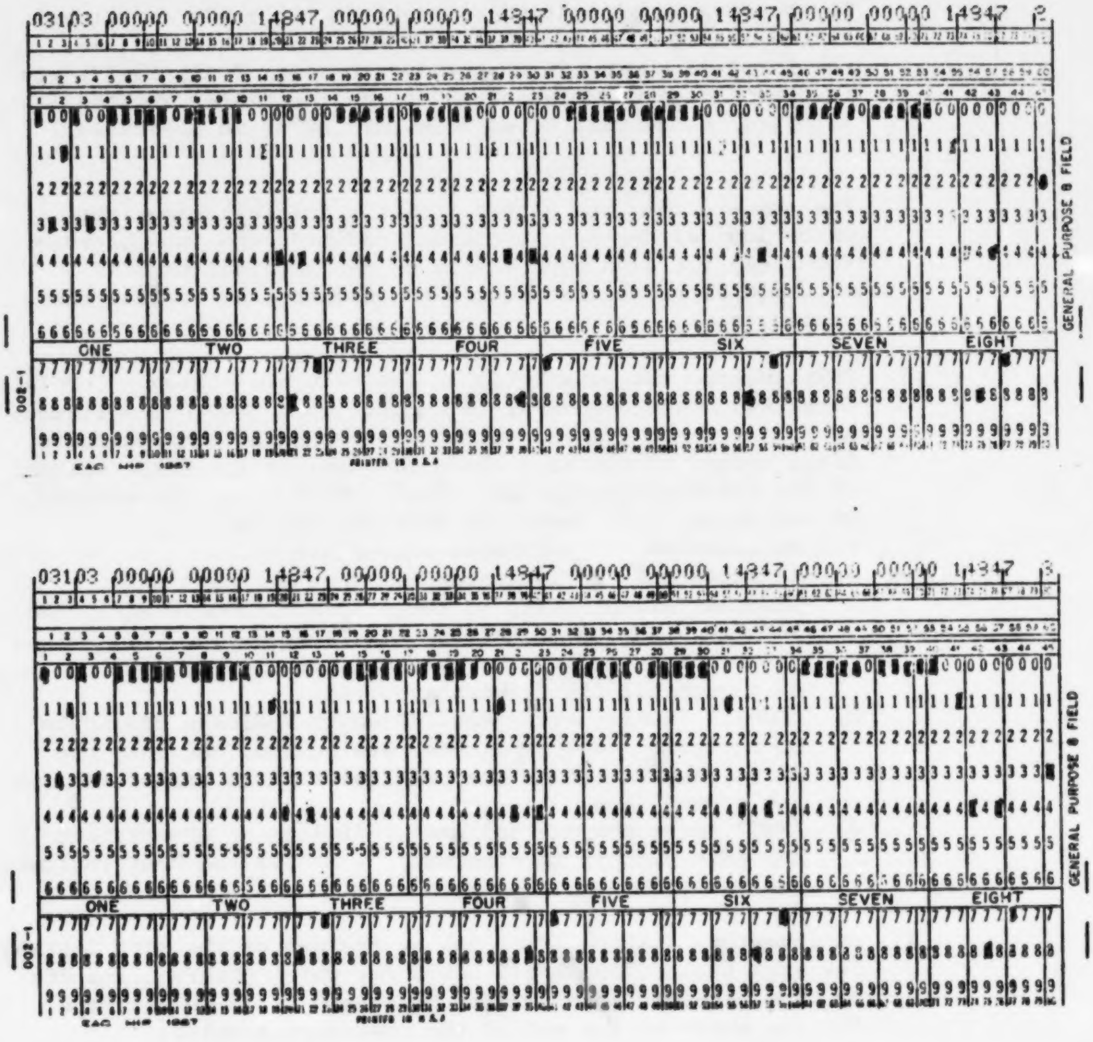

03103 00000 00000 14847 00000 00000 14847 00000 00000 14847 00000 00000 14847 2
ONE
TWO
THREE
FOUR
FIVE
SIX
SEVEN
EIGHT
GENERAL PURPOSE 8 FIELD
I-200
EAC MHP 1967
PRINTED IN U.S.A.
03103 00000 00000 14847 00000 00000 14847 00000 00000 14847 00000 00000 14847 3
ONE
TWO
THREE
FOUR
FIVE
SIX
SEVEN
EIGHT
GENERAL PURPOSE 8 FIELD
I-200
EAC MHP 1967
PRINTED IN U.S.A.

Fig. 1

5. Corrections for updating the tape are keyed on a "correction card" image by fields or by keying a complete No. 1 card image for missing observations or those having numerous field errors. Following the updating of the tape, another edit is made including the inventory listing.

IV Details of No. 1 Card Image Edit

A. Major check groups
   1. Psychrometric check: relationships between T, T_w, T_{dp}, & RH.
   2. Limiting ranges of variability.
   3. Wind, weather, temperature, and visibility and certain interrelationships.
   4. Cloud coding.
   5. Cloud, ceiling, and sky cover relationships.

B. Psychrometric Check

Psychrometric relationships. The program is designed to accept and check the interrelationships between the four psychrometric parameters if all are keyed, or to compute T_w and RH and then check the interrelationships if only T and T_{dp} are keyed. The notations are in terms of whole degrees Fahrenheit for temperature and whole percentages for relative humidity. If there is a suspected error in these relationships, the observation is printed out complete, including an appropriate error flag.

The empirical formulas used to compute T_w and RH (with respect to water) are:

1. Computation of wet bulb (T_w):
   If the dry-bulb temperature (T) is zero and above:

$$T_w = T - (.034N - .00072N \left[N - 1\right]) (T + T_{dp} - 2P + 108)$$

   a. If T is less than 100°F., rounding of T_w follows this scheme:

      T_w rounded = $T_w + .9$ if the tens position of T is 0, 1, 2.

      T_w rounded = $T_w + .9 - .01(T + .9)$ if tens position of T is 3, 4.

      T_w rounded = $T_w + .4$ if the tens position of T is 5 thru 9.

   b. If T is 100°F. or higher:

      T_w rounded = $T_w + .9$.

If the dry-bulb temperature (T) is less than zero:

$$T_w = T - (.034N - .006N^2)(.6\left[T + T_{dp}\right] - 2P + 108)$$

$$T_w \text{ rounded} = T_w - .01T_{dp}$$

$N = \frac{T - T_{dp}}{10}$ in the above equations.

2. Computation of relative humidity:

$$RH \approx \left(\frac{173 - .1T + T_{dp}}{173 + .9T}\right)^8$$

The checking procedures print out the error flag if:

T_{dp} is greater than T_w and if T_w is greater than T, and if the following are not satisfied:

a. In the range of temperature from -60° to +139°, the dew point range may be -60° to +90°. For individual observations, the dew point check requires that a maximum T_{dp} taking $T - 0.5°F.$ and $T_w + 0.4°F.$, and a minimum T_{dp} taking $T + 0.4°F.$, and $T_w - 0.5°F.$ (If $T = T_w$, maximum $T_{dp} = T_w$.) Saturation vapor pressures from tables stored in memory: Table A (vapor pressure over water for the range -60° to +140°F.) or B (vapor pressure over ice for the range -60° to +31°F.) are taken for the above values of T_w. The vapor pressures for each end of the allowable dew point range are then computed, using

$$e = e_w - 0.000367P\,(T - T_w)\left(1 + \frac{T_w - 32}{1571}\right)$$

(e, e_w, and P are in inches of mercury. Pressure may be taken from individual observation, or from the pressure applicable to elevation range in which station is located.)

From the vapor pressure tables in memory, the dew point temperatures corresponding to the vapor pressures at each end of the range, which are for the air at temperatures $T + 0.4°F.$, $T_w - 0.5°F.$, and $T - 0.5°F.$, $T_w + 0.4°F.$, are taken in terms of whole degrees of dew point. If the dew point being checked falls with 2° above or 2° below, it is accepted as correct. If outside this range, an indication of psychrometric error is printed. Note that if station pressure values are not recorded in the observations, computation of T_w should still always be possible since the program will taken an appropriate pressure value that corresponds to the station elevation.

b. Relative humidity values are accepted if they are in the range of 4% to 100% and are within 2% above and 2% below the computed range of humidity below. All values less than 4% are flagged for review. For hygrothermometer stations, the relative humidity will have been computed by the formula in 2 above; for other stations it will have been keyed from the original record.

The range for relative humidity is determined in the same way as for the dew point check. Maximum and minimum vapor pressures are obtained from the taped tables for each end of the range, and the computation at each end of the range is by this formula:

$RH = \frac{e}{e_s}$, e being the vapor pressure of the dew points and e_s the saturation vapor pressure of the air at the observed temperature plus 0.4°F or less 0.5°F.

If liquid fog is reported in present weather and the temperature is 31°F. or less, $T_{dp} = T_w = T$ is acceptable.

If T is less than -35°F., no formula is applied. In the latter case, when T = -36° or - 37°, an error is listed if the dew point does not fall within the range T - 6° (plus of minus 1°). An error also lists if temperature is within the range -38° through -53° and dew point is not in the range T - 7° (plus or minus 1°).

c. Limiting Range of Variability

Limiting values, some absolute and some dependent on other elements within an observation, are incorporated into the machine edit program for checking purposes. Items with values outside the limits, or such as appear inconsistent with other elements in the observation, or approach extreme conditions are flagged for technical review as follows:

1. Sea-level pressure: above 1060.0 or below 940.0 mb.
2. Station pressure: if pressure in inches and hundredths plus 10^{-3} times the elevation (H_p) in feet is less than 27.75 or greater than 31.30 inches.
3. Change of sea-level pressure from one observation to the next is greater than 6.0 mb., change of station pressure from one observation to the next is greater than 0.20 inches. The interval between observations in both cases is 3 hours. For 1-hour, 3.0 mbs. & 0.10 inch apply.
4. Temperature: T, above 125° or below -60°; T_w, above 125° or below -60°; T_{dp}, above 90° or below -60°, and if T_w and T_{dp} are present and T is -53° or colder.

5. Temperature fluctuation from one 3-nourly observation to the next: if T or T_{dp} changes 20° or more from one 3-hourly observation to the next, the observation which varies 20° or more from the preceding is flagged for review. Changes of 10° are flagged for hourly observations.

6. Relative humidity: below 4%.

7. Winds: When wind at one 3-hourly observation of 20 knots or more doubles at the next 3-hourly observation, or reaches 50 knots, the wind speed is flagged for review. (In AWS version of this edit, all winds 30 knots and higher are flagged for review.)

8. Visibility: is 15 miles or less at one observation and 70 or above at the next.

9. Obscuration and cloud heights, as follows:

a.	Obscuration	greater than 4,000 ft.
b.	Fog	greater than 1,500 ft.
c.	Stratus, stratocumulus, stratus fractus, cumulus fractus, cumulus mamatus	greater than 9,000 ft.
d.	Cumulus, cumulonimbus	greater than 12,000 ft.
e.	Altostratus, altocumulus, nimbostratus, and altocumulus castellanus	less than 4,500 ft. and greater than 20,000 ft.
f.	Cirrus, cirrostratus, and cirrocumulus	less than 15,000 ft.

D. Wind, Weather, Temperature, Visibility:

1. Wind: direction is recorded and keyed in tens of degrees from north (00 = Calm), and wind speed in knots (00 = Calm). If speed is 00, direction must be 00. Legal directions other than 00 are 01 through 36. The wind error indication is printed with illegal directions, for speed of 01 or more with direction 00, for direction of 01 - 36 with speed 00, and for exceeding limits mentioned in b above.

   Speed is related to the check of blowing dust, sand, blowing spray and blowing snow. Observations in which these items appear with wind speed less than 9 knots are flagged.

2. Weather: the following items and observations containing them are flagged for review:

   a. Tornado.

   b. Ice crystals with intensity indication or in combination with any other element.

c. Fog or any form of precipitation with clear sky (0 cloud amount) except ice crystals.

d. Fog with dew point depression greater than 8°F.

e. Fog with less than 1/10 cloud cover.

f. Weather types below with visibilities other than those listed:

Weather	Visibility range
S+, SP+,SW+, L+, ZL+, SG+	000-004 (0 - 1/4 mile)
S, SP, SW, L, ZL, SG, IC*	005-007 (5/16 - 1/2 mile)

(* Note: IC may be reported with higher than 1/2 mile visibility)

S-, SP-, SW-, L-, ZL-, SG-	008---- (3/4-unlimited)
F, IF, GF, BD, BN, K, H,KH, D, BS, BY	000-060 (0 to 6 miles)

g. Weather types (all intensities) with temperatures other than within ranges below:

Weather	Range of temperature
R, RW, L	28°F. or higher
ZR, ZL	No lower limit, to 39°F.
IP	10°F, through 44°F.
SP, SG, S, SW	-40°F. through 44°F.
IC	-40°F. through 15°F.
IF	-40°F. through 15°F.

h. 100% relative humidity reported without liquid fog or liquid precipitation in the weather fields and wind speed ≥6 knots.

i. Illegal visibility codes are flagged for correction. The legal visibility codes are:

VSBY	Code	VSBY	Code	VSBY	Code
0	000	1 5/8	018	8	080
1/16	001	1 3/4	019	9	090
1/8	002	2	020	10	100
3/16	003	2 1/4	024	11	110
1/4	004	2 1/2	027	12	120
5/16	005	3	030	13	130
3/8	006	4	040	14	140
1/2	007	5	050	15	150
5/8	008	6	060	20	200
3/4	009	7	070		
1	010			and, by 5 mile increments, on to	
1 1/8	012				
1 1/4	014				
1 3/8	016			95	950
1 1/2	017			≥100	990

E. Cloud Coding

Ceiling, sky condition, and clouds are interrelated. Three columns are keyed for ceiling height. The valid codes are as indicated below and any others are flagged for correction.

Ceiling height	Card code
Unlimited	XXX
Zero	000
100 ft. - 5000 ft. (every hundred feet)	001 - 050
5000 ft. - 10,000 ft. (every five hundred feet)	050 - 100
10,000 ft. and higher (every thousand feet)	100 - 250, etc.

Sky condition is a four-position (4 card columns) field, with provision for keying four sky condition symbols, as may be recorded in the MF1-10A Sky column. Heights of clouds are not keyed in this field (ceiling is keyed in the ceiling field and cloud heights in the "layer" fields discussed below). If less than 4 symbols are reported, keying begins at the left of the field, with "0" keyed in each column at the right of the field for which no sky symbol is reported. The lowest sky symbol is keyed first, the next highest second, etc., until the 4-column field is coded completely, either with sky condition symbols (including blanks) or zeros.

If more than 4 sky condition symbols are reported, the highest is keyed in column 20, and the first three in columns 17-19, unless this excludes the ceiling symbol. In the latter case the ceiling symbol is keyed in column 19, the first two in columns 17 and 18, and the highest in column 20.

For a partial obscuration (-X) the first column of the sky condition field is left blank. The succeeding three columns are keyed for reported sky conditions.

No clouds or obscurations (clear) is keyed 0000.

An obscuration (not partial) requires an X key in the first or second column of the sky condition field. If the obscuration is the lowest sky condition, the X will be in the first column. If a cloud layer is reported below the obscuration it will be keyed in the first column in the normal manner, and the X in the second column of the field. In this situation, the last two columns of the field would be 00.

The table below presents the valid codes of the Sky Condition field. In the table, p = punch, b = blank, and - = X.

Card code	Card column punching possibilities				Description
0000	0	0	0	0	Clear sky, (less than 1/10).
p000	1,2,4 5,7,8	0	0	0	One symbol only, not an obscuration or partial obscuration.
pp00	1,2,4 5,7	1,2,4 5,7,8	0	0	Two symbols reported, no obscuration or partial obscuration.
ppp0	1,2,4 5,7	1,2,4 5,7	1,2,4 5,7,8	0	Three symbols reported, no obscuration or partial obscuration.
pppp	1,2,4 5,7	1,2,4 5,7	1,2,4 5,7	1,2,4 5,7,8	Four symbols, no obscuration or partial obscuration
-000	X	0	0	0	Obscuration, 10/10 sky hidden, no layer below obscuration.
b000	Blank	0	0	0	Partial obscuration, no other symbols.
bp00	Blank	1,2,4 5,7,8	0	0	Partial obscuration, one other symbol.
bpp0	Blank	1,2,4 5,7	1,2,4 5,7,8	0	Partial obscuration, two other symbols.
bppp	Blank	1,2,4 5,7	1,2,4 5,7	1,2,4 5,7,8	Partial obscuration, three other symbols.
p-00	1,2,4	X	0	0	Obscuration above one layer of cloud

F. Clouds and Obscuring Phenomena.

Provision for keying as many as four layers of clouds and/or obscuring phenomena, total sky cover, and opaque sky cover amount is made in this field. Cloud layers are keyed in ascending order. If more than four layers are reported, the four lowest are keyed. The lowest layer is always keyed in the left hand cloud field of the card. For each layer, amount, type, and height are keyed. For the second and third layers (if reported), the summation amount(s) is keyed at the level(s) involved.

If a complete cloud layer section is reported unknown, "U", on MF1-10B, the corresponding card field for the entire layer is left blank.

When fog or any other obscuring phenomenon is reported, it will be handled in a manner similar to a cloud layer, and an amount, type, and height will be keyed. Obscuring phenomena other than fog (smoke, for example) are keyed X for type. Heights of clouds and vertical visibility into obscurations are keyed in hundreds of feet. Where vertical visibility is unlimited (dash in height column of MF1-10B) height is keyed XXX. If cloud field is reported clear or none, height will be keyed XXX. If cloud height is reported unknown (U), height is left blank if type is unknown.

Summation totals may not exceed 10/10, but the first summation (card col. 67) may be 1 greater (not exceeding 10/10) than the sum of card columns 57 and 62; and card column 73 may be 1 greater than the sum of card columns 67 and 68, not to exceed 10/10. Total cloud amount (card col. 56) should be the same as col. 57 if only one layer is reported, the same as col. 67 if only two layers are reported, the same as col. 73 if only three layers are reported, and equal to not more than 1 greater than the sum of cols. 73 and 74 (not exceeding 10/10) if four layers are reported.

1. Legal codes in the card field for "Clouds and Obscuring Phenomena" are related to the Ceiling, Sky Condition, and Weather and/or Obstruction to Vision fields. Accordingly, a discussion of the several relationships is presented.

   a. If sky condition is reported clear, ceiling must be unlimited. Summation of all clouds must be zero. Type and height in the cloud layer fields may be keyed for zero amount (less than 1/10).

b. The ceiling height must be consistent with the height of the lowest cloud layer whose corresponding symbol in the sky condition field is broken or overcast, or with the height of an obscuration. The total (if one layer) or the summation amount at the layer constituting the ceiling must be equal or greater than 6/10. The cloud type must be coded either 2 through 9, X/2, X/4 through X/7, X/9, or, if an obscuration, 1 or X.

c. If the ceiling is not XXX (unlimited), some sky symbol must be keyed: i. e., broken (5), ovc (8), or obscured (X in 1st or 2nd col. of sky condition field). Only one X may be keyed. It will be the first column of the sky condition field if the lowest layer is the obscuration; the second is a layer below a portion of the surface-based obscuration.

d. If the first cloud layer contains 10/10 F or IF (not GF), the ceiling height must equal the height of the first layer, and the sky must be obscured.

e. If fog is keyed as an Obstruction to Vision with clear sky or with partial obscuration or less than 5/10, with no clouds above, the fog must be GF or IF (not F). If the partial obscuration is 6/10 or greater, with no clouds above, or with obscuration (10/10), the fog may not be classified as GF.

f. If total opaque is zero, all sky symbols must be thin or clear, and ceiling must be unlimited.

g. If any sky symbol is thin, the total opaque amount must not be more than half the summation amount of that layer and all higher layers (not always in error for higher layers, but should be flagged for review). If the ratio of total amount is 1/2 or less, the highest sky symbol must be thin.

h. Sky condition symbols must, with increasing height, reflect equal or increasing sky cover. Only the highest sky symbol may be overcast, except that below an overcast there may be a thin overcast.

i. The highest sky condition symbol must be compatible with the amount of total sky cover.

j. If obscuration (X) is reported as the second sky condition, the second cloud layer type must be obscuring phenomena (fog, ice fog, smoke, rain, snow, for example) keyed X; total amount, total opaque, and first summation total must be 10/10. The third and fourth cloud layers and the second summation total columns should be blank (may be keyed if an aircraft report has been received, but should be flagged for review). The third and fourth sky cover symbols must be zero (0). Ceiling must not exceed the second cloud layer height, and that height should be 4,000 ft. or less. Normally fog will be questioned in such a situation.

k. If obscuration is reported as the first sky symbol, the type of the first cloud layer must be fog (code 1) or other obscuring phenomena (code X), total amount and total opaque must be 10/10, and height must correspond with ceiling height. The second, third, and fourth cloud layer and first and second summation total columns should be blank (may be keyed if an aircraft report has been received, but should be flagged for review). The second, third, and fourth sky cover symbols must be zero (0). Height and ceiling should be the same. Height should be 1,500 ft. or less if fog (code 1) or 4,000 ft. or less if other obscurations (X) are encoded.

l. When fog (code 1) is reported in the first cloud layer, amount not coded 0, the sky condition must reflect an obscuration or partial obscuration.

m. When fog is reported as the only cloud field (code 1), it should be coded in Obstructions to Vision as GF if amount is 1 to 5 tenths, or F if 5 to 10 tenths (prevailing visibility being 6 miles or less).

n. The corresponding cloud and summation total columns for sky cover symbol reported (code) above an overcast (code 8 in Sky Condition) should be:

-1. Blank if total opaque is 10/10.
-2. Zero (0) in amount and type columns, 10 in summation total columns, and XXX in height columns whenever total opaque is less than 10/10. Additional layers may be keyed if an aircraft report has been received, but should be flagged for review).

o. Partial obscuration (blank in first position of Sky Condition) must have a first-layer amount from 1 to 9, type must be fog (code 1) or obscuration (code X), and height must be unlimited.

p. Some stations (FAA) do not observe cloud layer values, but do enter total cloudiness and total opaque. If the ratio of total opaque to total amount is 1/2 or less, there should be no codes 5, 8, or obscuration (X) in Sky Condition. If the ratio is greater than 1/2, there must be 5, 8, or X in Sky Condition, and if ratio is not 1:1, X is invalid. (The valid blanks in cloud layer field will cause "2's" to print in the inventory listing for these stations.

2. The testing procedure to flag errors or suspected conditions in Clouds and Obscuring Phenomena, Sky Condition, Total Clouds, and Total Opaque is systematic. Missing fields are indicated (except the valid condition for FAA coded "2" in the inventory) in the usual manner. The system, in general follows these steps:

In the cloud fields, the valid codes for cloud amounts are 0 = no clouds or less than 1/10; 1 - 9 = 1/10 to 9/10 clouds; X = more than 9/10 or 10/10.

Valid codes for cloud types are 0 = NONE; b (Blank) when a cloud type is reported UNKNOWN; 1 = Fog; 2 = Stratus; 3 = Stratocumulus; 4 = Cumulus; 5 = Cumulonimbus; X/2 (K) = Stratus fractus; X/4 (M) = Cumulus fractus; X/5 (N) = Cumulus mammatus; 6 = Altostratus; 7 = Altocumulus; X/6 (O) = Nimbostratus; X/7 (P) = Altocumulus castellanus; 8 = Cirrus; 9 = Cirrostratus; X/9 (R) = Cirrocumulus; X = Obscuration other than fog.

The valid codes for cloud heights in the cloud layer fields are the same as for ceiling heights (in number of hundreds of feet); XXX indicates NONE or (in the first layer) a surface-based partial obscuration; and bbb (Blanks) indicates cloud height unknown with type unknown.

Errors are listed for invalid codes, and if

a. Any cloud field element is keyed and the Total Clouds left blank.
b. Total opaque is not keyed, unless indicated in the station's observation pattern (III. D. 1.).
c. Total opaque is greater than total cloud amount.
d. Total opaque is less than 10/10, and any blanks occur in Cloud and Obscuring Phenomena fields including summarions, FAA excepted.
e. Any element within a cloud layer is keyed (amount, type, or height), and any other element is left blank.
f. Total Cloud amount is keyed from 0/10 to 9/10 inclusive, and amounts and types of fields above highest reported layer are not coded "O" and heights are not coded "XXX."
g. Each summation amount does not equal or exceed the next lower summation amount, or if a succeeding summation amount is greater than 1 more than the amount(s) of the accitional layer(s), or exceeds 10/10.
    1. In the case of partial obscuration in the first layer, the second summation is not greater than the amount of the first layer if a cloud layer is also reported.

2. The summation amount is less than any lower individual cloud layer amount.

3. Blanks in a summation amount are not preceded by 10/10 in the last summation amount (which is not blank) by blank cloud amounts in fields with blank summarion amounts, and 10/10 Total Opaque is not reported.

h. Height ranges by cloud types are in disagreement with those listed under Obscuration and Cloud heights.

i. Fog (code 1) is coded in layers above the first.

j. Height in cloud layers reporting height does not increase from one layer to the next.

k. The ceiling height does not agree with the lowest layer height constituting a ceiling, or the highest sky symbol is not compatible with total sky cover.

G. Explanation of Edit Flags

Reproduced from best available copy.

Beginning on the following page are numbered explanations of edit flags appearing in the correspondingly numbered observations printed out in the Surface Weather Observations format. An asterisk (*) prints over fields to be checked.

The edit is designed for No. 1 card images keyed every 3 hours (at local standard time corresponding to 00, 03, 06, 09, 12, 15, 18 and 21 hours GMT) and is readily adaptable to programs under which all record observations are keyed.

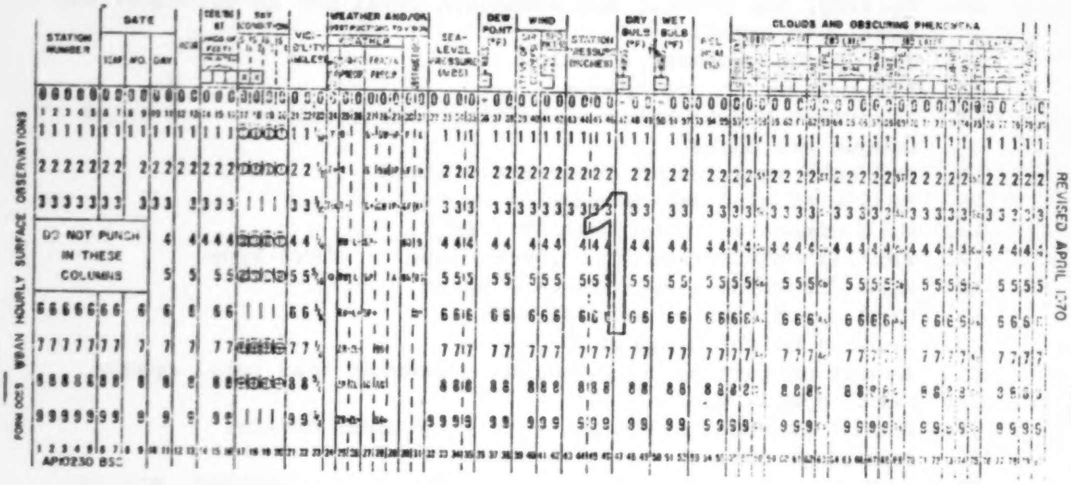

STATION NUMBER
DATE
YEAR
MO.
DAY
HOUR
CEILING
SKY CONDITION
VISIBILITY
WEATHER AND/OR OBSTRUCTIONS TO VISION
SEA-LEVEL PRESSURE
DEW POINT
WIND
STATION PRESSURE (INCHES)
DRY BULB (°F)
WET BULB (°F)
REL. HUM. (%)
CLOUDS AND OBSCURING PHENOMENA
DO NOT PUNCH IN THESE COLUMNS
1
FORM 0059 WBAN HOURLY SURFACE OBSERVATIONS
REVISED APRIL 1970
AP10230 BSC

Fig. 2

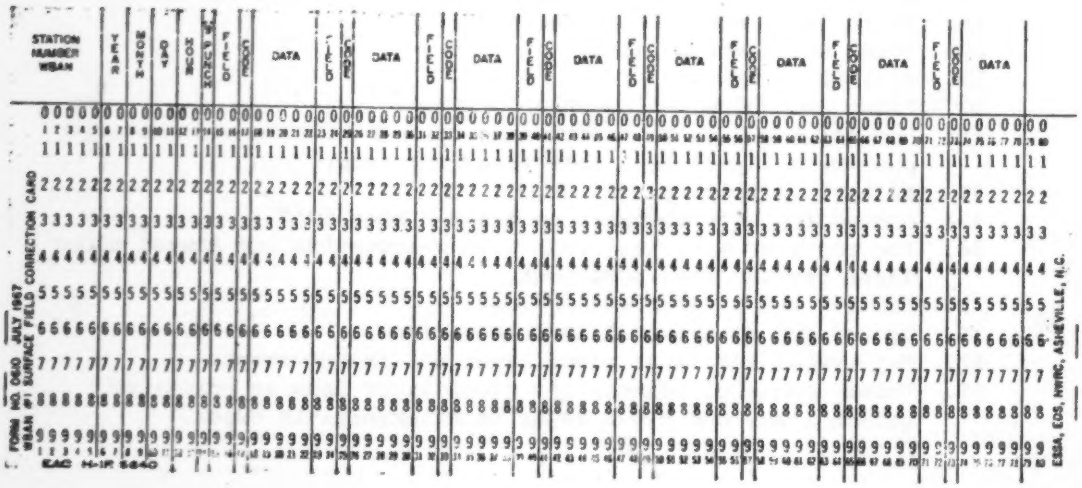

STATION NUMBER WBAN
YEAR
MONTH
DAY
HOUR
NUMBER
FIELD
CODE
DATA
FORM NO. 0560 JULY 1967
WBAN #1 SURFACE FIELD CORRECTION CARD
ESSA, EDS, NWRC, ASHEVILLE, N.C.

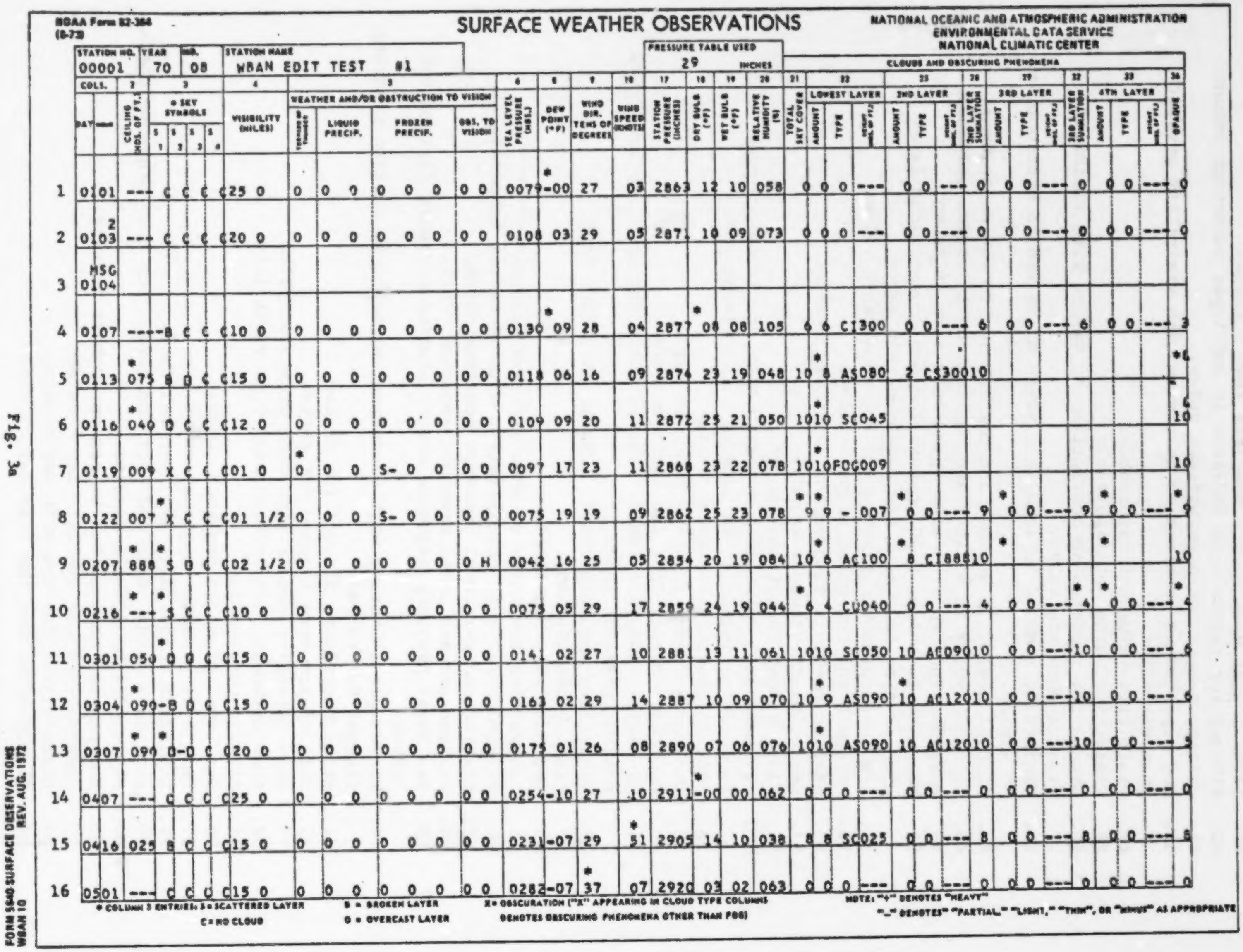
Fig. 2a

Explanations of Reasons for Flags on Correspondingly Numbered Observations in the Edit Listings **

1. Dew point incorrectly keyed.
2. Hour other than that in the normal program.
3. Card missing for hour in station program.
4. Dew point temperature higher than dry-bulb temperature.
5. Ceiling height differs from cloud layer height.
6. Ceiling height differs from cloud layer height.
7. Obscuration under Weather is snow with type of obscuration fog.
8. Obscuration with less than 10/10 sky cover.
9. First sky symbol scattered with corresponding layer over 5/10, and Second cloud group and ceiling non reportable value.
10. Ceiling, sky symbol, and summation total not in agreement with total sky cover.
11. Two opaque overcast symbols.
12. Lower layer is not opaque.
13. Incorrect relationship of ceiling, sky condition, and total opaque sky cover.
14. Dry-bulb temperature incorrectly keyed.
15. Review of wind speeds over 50 knots.
16. Wind direction value over 36 (360°).
17. Partial obscuration with lowest layer a cloud type.
18. Wind direction with calm wind speed.
19. Wind speed with no wind direction.
20. Ceiling height missing.
21. Amount of partial obscuration is greater than total opaque.
22. Total opaque missing.
23. Amount of obscuration and total sky cover differ.
24. Amount of obscuration less than 10/10.
25. Partial obscuration due to fog omitted from weather, and incomplete keying of cloud and obscuring phenomena.
26. Amount of partial obscuration is greater than total opaque. Also, incomplete keying of clouds and obscuring phenomena.
27. Third layer summation is missing.
28. Total sky cover omitted.
29. Visibility omitted.
30. First two columns of weather omitted.
31. Fog not shown as obstruction to vision. Clouds and obscuring phenomena layers less than total sky cover.
32. Illegal visibility.
33. Ground fog with obscuration greater than 5/10.
34. Fog reported with less than 6/10 obscuration and lowest cloud layer greater than 5000 feet.
35. Ceiling not a reportable height.
36. Ground fog with over 5/10 obscuration.
37. Sky symbol and first cloud layer not in agreement.
38. Total opaque cloudiness and cloud layer data in error; or ceiling, sky and cloud layer relationships in error.

** See pages 168 through 174.

39. Blowing dust with wind speed less than 7 knots.
40. Flagged for review - no increase in 2nd layer summation amount.
41. Ceiling height not a reportable value.
42. Fog reported as obstruction to vision with visibility greater than six miles.
43. Visibility value not reportable.
44. Visibility value not reportable.
45. Ground fog reported as obstruction to vision with visibility greater than six miles.
46. Visibility reduced to less than seven miles and no obstruction to vision.
47. Illegal keying in weather and obstruction to vision columns.
48. Illegal keying in weather and obstruction to vision columns.
49. Squalls reported with wind speed less than 16 knots.
50. Fog with less than 6/10 obscuration and lowest cloud layer greater than 5000 feet and psychrometric error.
51. Sea level pressure flagged for non-reportable value.
52. Dry bulb and dew point sequence check.
53. Station pressure flagged for improbable value.
54. Dew point incorrectly keyed.
55. Cloud height incorrectly keyed.
56. Illegal punch in weather & obstruction to vision columns.
57. Flagged for intensity of snow with 1/4 visibility.
58. Station pressure sequence check.
59. Duplicate cards, date and hour 1st card.
60. Duplicate cards, date and hour 2nd card.
61. Visibility sequence check. Change in values up. Sea level and station pressure check.
62. Visibility sequence check. Change in values down.
63. Missinb observation.
64. Dry bulb sequence check.
65. Flagged for intensity of snow with 1/4 mile visibility.
66. Flagged for liquid precipitation with 24-degree temperature.
67. Station and sea level pressure sequence check.
68. Station and sea level pressure sequence check.
69. Sea level pressure sequence check, station pressure flagged for review.
70. Frozen precipitation with 45-degree temperature, station pressure flagged for review.
71. Station pressure flagged for review.
72. Station pressure flagged for review.
73. Snow intensity not in agreement with visibility.
74. Snow intensity not in agreement with visibility.
75. Missing observation.
76. Sea level pressure sequence check.
77. Sea level pressure sequence check.
78. Duplicate cards, date and hour 1st card.
79. Duplicate cards, date and hour 2nd card. Dry bulb sequence check.
80. Sky condition symbols missing.
81. Weather and obstruction to vision symbols missing.
82. Observations for the 29th day missing.
83. Observations for the 30th day missing.
84 Observations for the 31st day missing.
85. Monthly inventory check.



NOAA Form 82-384
(8-73)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

STATION NO.	YEAR	MO.	STATION NAME
00001	70	08	WBAN EDIT TEST #1

PRESSURE TABLE USED 29 INCHES

	COLS.		2	3				4	5				6	8	9	10	17	18	19	20	21	22			25			26	29			32	33			36
				* SKY SYMBOLS					WEATHER AND/OR OBSTRUCTION TO VISION												CLOUDS AND OBSCURING PHENOMENA															
																						LOWEST LAYER			2ND LAYER				3RD LAYER				4TH LAYER			
	DAY	HOUR	CEILING (HDS. OF FT.)	S 1	S 2	S 3	S 4	VISIBILITY (MILES)	TORNADO & THUNDER	LIQUID PRECIP.	FROZEN PRECIP.	OBS. TO VISION	SEA LEVEL PRESSURE (MBS.)	DEW POINT (°F)	WIND DIR. TENS OF DEGREES	WIND SPEED (KNOTS)	STATION PRESSURE (INCHES)	DRY BULB (°F)	WET BULB (°F)	RELATIVE HUMIDITY (%)	TOTAL SKY COVER	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	2ND LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	3RD LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	OPAQUE
1	01	01	---	C	C	C	C	25 0	0	0 0	0 0 0	0 0	0079	*-00	27	03	2863	12	10	058	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
2	01	Z 03	---	C	C	C	C	20 0	0	0 0	0 0 0	0 0	0108	03	29	05	2871	10	09	073	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
3	01	MSG 04																																		
4	01	07	---	-B	C	C	C	10 0	0	0 0	0 0 0	0 0	0130	*09	28	04	2877	*08	08	105	6	6	CI	300	0	0	---	6	0	0	---	6	0	0	---	3
5	01	13	*075	B	O	C	C	15 0	0	0 0	0 0 0	0 0	0118	06	16	09	2874	23	19	048	10	*8	AS	080	2	CS	300	10								*6
6	01	16	*040	O	C	C	C	12 0	0	0 0	0 0 0	0 0	0109	09	20	11	2872	25	21	050	10	*10	SC	045												6 10
7	01	19	009	X	C	C	C	01 0	*0	0 0	S- 0 0	0 0	0097	17	23	11	2868	23	22	078	10	*10	FOG	009												10
8	01	22	007	*X	C	C	C	01 1/2	0	0 0	S- 0 0	0 0	0075	19	19	09	2862	25	23	078	*9	*9	-	007	*0	0	---	9	*0	0	---	9	*0	0	---	*9
9	02	07	*888	*S	O	C	C	02 1/2	0	0 0	0 0 0	0 H	0042	16	25	05	2854	20	19	084	10	*6	AC	100	*8	CI	888	10	*				*			10
10	02	16	*---	*S	C	C	C	10 0	0	0 0	0 0 0	0 0	0075	05	29	17	2859	24	19	044	*6	4	CU	040	0	0	---	4	0	0	---	*4	*0	0	---	*4
11	03	01	050	*O	O	C	C	15 0	0	0 0	0 0 0	0 0	0141	02	27	10	2881	13	11	061	10	10	SC	050	10	AC	090	10	0	0	---	10	0	0	---	6
12	03	04	*090	-B	O	C	C	15 0	0	0 0	0 0 0	0 0	0163	02	29	14	2887	10	09	070	10	*9	AS	090	*10	AC	120	10	0	0	---	10	0	0	---	6
13	03	07	*090	*O	-O	C	C	20 0	0	0 0	0 0 0	0 0	0175	01	26	08	2890	07	06	076	10	*10	AS	090	10	AC	120	10	0	0	---	10	0	0	---	5
14	04	07	---	C	C	C	C	25 0	0	0 0	0 0 0	0 0	0254	-10	27	10	2911	*-00	00	062	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
15	04	16	025	B	C	C	C	15 0	0	0 0	0 0 0	0 0	0231	-07	29	*51	2905	14	10	038	8	8	SC	025	0	0	---	8	0	0	---	8	0	0	---	5
16	05	01	---	C	C	C	C	15 0	0	0 0	0 0 0	0 0	0282	-07	*37	07	2920	03	02	063	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0

* COLUMN 3 ENTRIES: S = SCATTERED LAYER B = BROKEN LAYER X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS

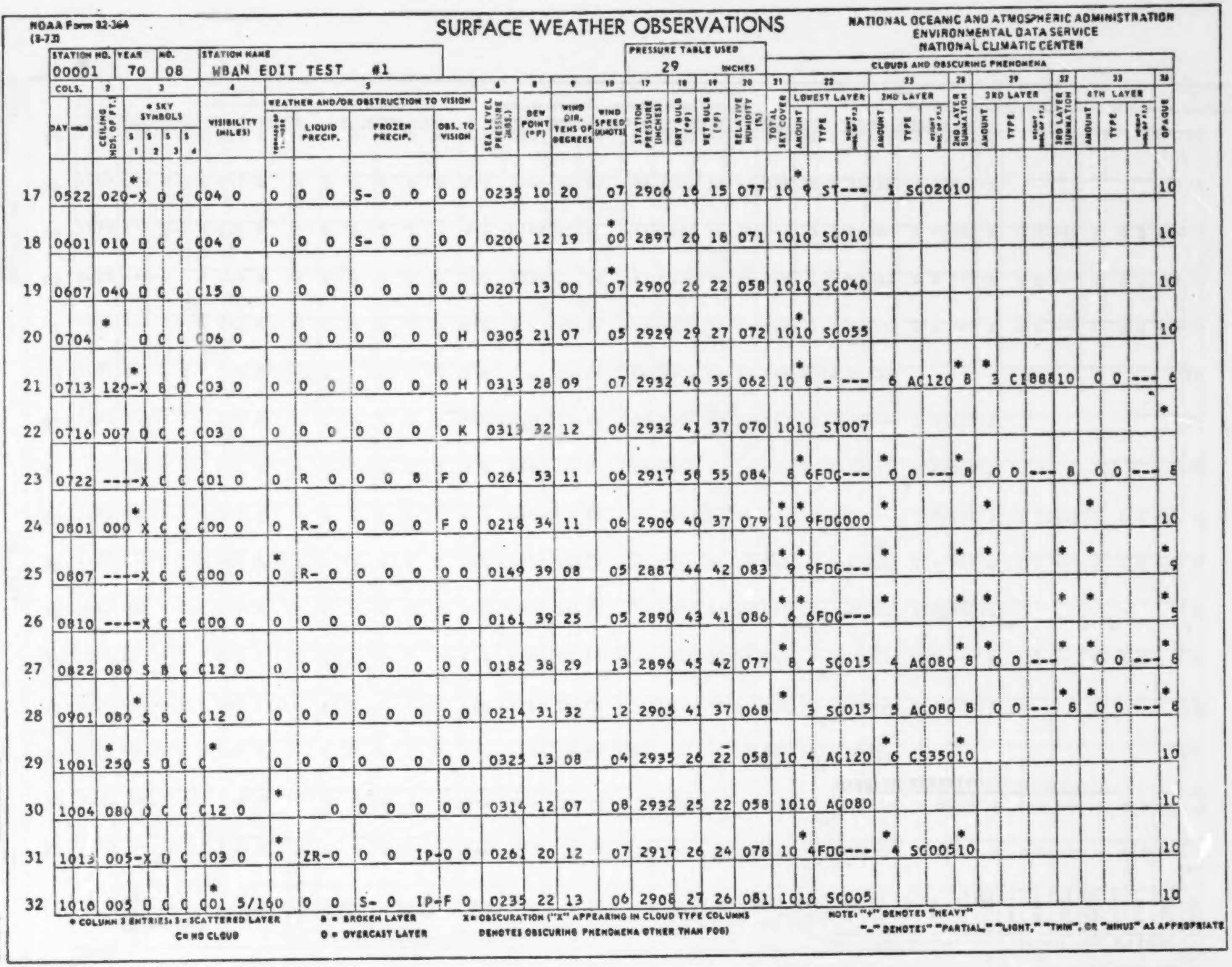
C = NO CLOUD O = OVERCAST LAYER DENOTES OBSCURING PHENOMENA OTHER THAN FOG)

NOTE: "+" DENOTES "HEAVY"

"-" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
REV. AUG. 1972
WBAN 10

Fig. 3a



NOAA Form 82-364
(3-72)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

STATION NO. 00001 | YEAR 70 | NO. 08 | STATION NAME WBAN EDIT TEST #1

PRESSURE TABLE USED 29 INCHES

CLOUDS AND OBSCURING PHENOMENA

	COLS.		2	3				4	5				6	8	9	10	17	18	19	20	21	22			25			28	29			32	33			36
				* SKY SYMBOLS					WEATHER AND/OR OBSTRUCTION TO VISION													LOWEST LAYER			2ND LAYER				3RD LAYER				4TH LAYER			
	DAY	[illegible]	CEILING HDS. OF FT.	S 1	S 2	S 3	S 4	VISIBILITY (MILES)	TORNADO [illegible]	LIQUID PRECIP.	FROZEN PRECIP.	OBS. TO VISION	SEA LEVEL PRESSURE (MBS.)	DEW POINT (°F)	WIND DIR. TENS OF DEGREES	WIND SPEED (KNOTS)	STATION PRESSURE (INCHES)	DRY BULB (°F)	WET BULB (°F)	RELATIVE HUMIDITY (%)	TOTAL SKY COVER	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	2ND LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	3RD LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	OPAQUE
17	05	22	020	* -X	0	C	C	04 0	0	0 0	S- 0 0	0 0	0235	10	20	07	2906	16	15	077	10	* 9	ST	---	1	SC	020	10								10
18	06	01	010	0	C	C	C	04 0	0	0 0	S- 0 0	0 0	0200	12	19	* 00	2897	20	18	071	10	10	SC	010												10
19	06	07	040	0	C	C	C	15 0	0	0 0	0 0 0	0 0	0207	13	00	* 07	2900	26	22	058	10	10	SC	040												10
20	07	04	*	0	C	C	C	06 0	0	0 0	0 0 0	0 H	0305	21	07	05	2929	29	27	072	10	* 10	SC	055												10
21	07	13	120	* -X	B	0	C	03 0	0	0 0	0 0 0	0 H	0313	28	09	07	2932	40	35	062	10	* 8	-	---	6	AC	120	* 8	* 3	CI	888	10	0	0	---	6
22	07	16	007	0	C	C	C	03 0	0	0 0	0 0 0	0 K	0313	32	12	06	2932	41	37	070	10	10	ST	007												*
23	07	22	---	-X	C	C	C	01 0	0	R 0	0 0 8	F 0	0261	53	11	06	2917	58	55	084	8	* 6	FOG	---	* 0	0	---	* 8	0	0	---	8	0	0	---	8
24	08	01	000	* X	C	C	C	00 0	0	R- 0	0 0 0	F 0	0218	34	11	06	2906	40	37	079	* 10	* 9	FOG	000	*				*				*			10
25	08	07	---	-X	C	C	C	00 0	* 0	R- 0	0 0 0	0 0	0149	39	08	05	2887	44	42	083	* 9	* 9	FOG	---	*			*	*			*	*			* 9
26	08	10	---	-X	C	C	C	00 0	0	0 0	0 0 0	F 0	0161	39	25	05	2890	43	41	086	* 6	* 6	FOG	---	*			*	*			*	*			* 5
27	08	22	080	S	B	C	C	12 0	0	0 0	0 0 0	0 0	0182	38	29	13	2896	45	42	077	* 8	4	SC	015	4	AC	080	* 8	* 0	0	---	*	* 0	0	---	* 8
28	09	01	080	* S	B	C	C	12 0	0	0 0	0 0 0	0 0	0214	31	32	12	2905	41	37	068	*	3	SC	015	5	AC	080	8	0	0	---	* 8	* 0	0	---	* 8
29	10	01	* 250	S	0	C	C	*	0	0 0	0 0 0	0 0	0325	13	08	04	2935	26	$\overline{22}$	058	10	4	AC	120	* 6	CS	350	* 10								10
30	10	04	080	0	C	C	C	12 0	*	0	0 0 0	0 0	0314	12	07	08	2932	25	22	058	10	10	AC	080												10
31	10	13	005	-X	0	C	C	03 0	* 0	ZR-0	0 0 IP-	0 0	0261	20	12	07	2917	26	24	078	10	* 4	FOG	---	* 4	SC	005	* 10								10
32	10	10	005	0	C	C	C	* 01 5/16	0	0 0	S- 0 IP-	F 0	0235	22	13	06	2908	27	26	081	10	10	SC	005												10

* COLUMN 3 ENTRIES: S = SCATTERED LAYER   B = BROKEN LAYER   X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS
C = NO CLOUD   O = OVERCAST LAYER   DENOTES OBSCURING PHENOMENA OTHER THAN FOG)

NOTE: "+" DENOTES "HEAVY"
"-" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
WBAN 10   REV. AUG. 1972

Fig. 3b

NOAA Form 12-364
(8-72)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

STATION NO.	YEAR	MO.	STATION NAME	PRESSURE TABLE USED
00001	70	08	WBAN EDIT TEST #1	29 INCHES

	COLS.		2	3				4	5				6	8	9	10	17	18	19	20	21	22			25			28	29			32	33			36
				* SKY SYMBOLS					WEATHER AND/OR OBSTRUCTION TO VISION												CLOUDS AND OBSCURING PHENOMENA															
																						LOWEST LAYER			2ND LAYER				3RD LAYER				4TH LAYER			
	DAY	HOUR	CEILING HDS. OF FT.	S 1	S 2	S 3	S 4	VISIBILITY (MILES)	TORNADO & THUNDER	LIQUID PRECIP.	FROZEN PRECIP.	OBS. TO VISION	SEA LEVEL PRESSURE (MBS.)	DEW POINT (°F)	WIND DIR. TENS OF DEGREES	WIND SPEED (KNOTS)	STATION PRESSURE (INCHES)	DRY BULB (°F)	WET BULB (°F)	RELATIVE HUMIDITY (%)	TOTAL SKY COVER	AMOUNT	TYPE	HEIGHT HDS. OF FT.	AMOUNT	TYPE	HEIGHT HDS. OF FT.	2ND LAYER SUMMATION	AMOUNT	TYPE	HEIGHT HDS. OF FT.	3RD LAYER SUMMATION	AMOUNT	TYPE	HEIGHT HDS. OF FT.	OPAQUE
33	10	19	---	* -X	C	C	C	05 0	* 0	0 0	0 0 0	GFO	0212	24	11	06	2902	29	28	082	* 9	* 9	FOG	---	0	0	---	9	0	0	---	9	0	0	---	* 9
34	10	22	060	-X	B	C	C	01 0	* 0	0 0	0 0 0	F 0	0190	24	05	08	2896	28	27	085	8	* 4	FOG	---	4	SC	060	8	0	0	---	8	0	0	---	8
35	12	07	* 054	O	C	C	C	10 0	0	0 0	0 0 0	0 0	0176	23	08	04	2893	32	29	069	10	* 10	SC	054												10
36	12	10	003	-X	O	C	C	00 1/2	* 0	R- 0	0 0 0	GFO	0156	28	12	05	2887	33	31	082	10	* 7	FOG	---	3	ST	003	10								10
37	12	13	010	* B	C	C	C	00 0	0	R- 0	0 0 0	F 0	0106	33	13	04	2873	34	34	096	8	* 8	FOG	010	0	0	---	8	0	0	---	8	0	0	---	8
38	13	07	055	B	O	C	C	15 0	0	0 0	0 0 0	0 0	0094	28	28	14	2869	31	30	089	* 10	7	SC	055	3	CI	300	* 10	*			*	*			* 5
39	13	10	---	C	C	C	C	06 0	* 0	0 0	0 0 0	BDO	0115	29	28	* 05	2875	32	31	089	0	0	AC	100	0	0	---	0	0	0	---	0	0	0	---	0
40	14	07	020	-X	S	O	C	06 0	0	0 0	0 0 0	F H	0239	10	29	04	2909	13	13	088	10	* 3	FOG	---	2	ST	010	* 3	7	ST	020	10	0	0	---	7
41	14	13	* 153	B	O	C	C	07 0	0	0 0	0 0 0	0 0	0240	10	07	05	2909	20	18	065	10	* 6	AS	153	* 4	CS	888	10								10
42	14	22	070	-X	O	C	C	* 15 0	* 0	0 0	0 0 0	GFO	0240	17	09	03	2909	20	19	088	10	3	FOG	---	7	AS	070	10								10
43	15	01	022	O	C	C	C	* 02 9	0	0 0	S- 0 0	0 0	0233	15	04	04	2907	18	18	088	10	10	SC	022												10
44	15	07	---	S	C	C	C	* 17 0	0	0 0	0 0 0	0 0	0251	11	31	05	2912	14	14	058	1	1	AC	080	0	0	---	1	0	0	---	1	0	0	---	1
45	16	10	---	-B	C	C	C	* 08 0	* 0	0 0	0 0 0	GFO	0238	17	25	10	2910	34	28	050	8	8	CI	300	0	0	---	8	0	0	---	8	0	0	---	4
46	16	16	---	S	-S	C	C	* 06 0	* 0	0 0	0 0 0	0 0	0184	23	27	14	2895	43	35	045	3	1	CU	060	2	CI	300	3	0	0	---	3	0	0	---	1
47	18	04	014	B	O	C	C	15 0	* 0	0 0	0 0 8	0 0	0138	30	20	06	2882	38	35	073	10	7	SC	014	3	SC	045	10								10
48	18	07	012	B	O	C	C	10 0	* 0	0 0	0 0 6	0 0	0101	24	20	09	2871	29	28	082	10	8	SC	012	2	SC	035	10								10

* COLUMN 3 ENTRIES: S = SCATTERED LAYER   B = BROKEN LAYER   X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS
C = NO CLOUD   O = OVERCAST LAYER   DENOTES OBSCURING PHENOMENA OTHER THAN FOG)

NOTE: "+" DENOTES "HEAVY"
"–" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
WBAN 10   REV. AUG. 1972

Fig. 3c

NOAA Form 32-364
(8-73)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

STATION NO.	YEAR	MO.	STATION NAME	PRESSURE TABLE USED
00001	70	08	WBAN EDIT TEST #1	29 INCHES

	COLS.	2	3				4	5				6	8	9	10	17	18	19	20	21	22			25			28	29			32	33			34
			∗ SKY SYMBOLS					WEATHER AND/OR OBSTRUCTION TO VISION												CLOUDS AND OBSCURING PHENOMENA	LOWEST LAYER			2ND LAYER				3RD LAYER				4TH LAYER			
	DAY HOUR	CEILING (HDS. OF FT.)	S 1	S 2	S 3	S 4	VISIBILITY (MILES)	TORNADO & THUNDER	LIQUID PRECIP.	FROZEN PRECIP.	OBS. TO VISION	SEA LEVEL PRESSURE (MBS.)	DEW POINT (°F)	WIND DIR. TENS OF DEGREES	WIND SPEED (KNOTS)	STATION PRESSURE (INCHES)	DRY BULB (°F)	WET BULB (°F)	RELATIVE HUMIDITY (%)	TOTAL SKY COVER	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	2ND LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	3RD LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	OPAQUE
49	1810	009	S	O	C	C	01 1/2	* Q	0 0	S- 0 0	F 0	0077	30	22	* 02	2864	31	31	096	10	3	SC	004	7	SC	009	10								10
50	1816	---	-X	-S	C	C	05 0	* 0	0 0	0 0 0	F 0	0043	* 18	29	14	2855	* 35	29	050	4	* 2	FOG	---	2	CI	300	4	0	0	---	4	0	0	---	2
51	1901	---	S	C	C	C	15 0	0	0 0	0 0 0	0 0	* 0628	10	30	11	2850	22	19	060	1	1	AC	100	0	0	---	1	0	0	---	1	0	0	---	1
52	1916	---	S	-S	C	C	30 0	0	0 0	0 0 0	0 0	0009	002 -20	30	21	2843	021 -10	-10	060	4	2	SC	035	2	CI	350	4	0	0	---	4	0	0	---	2
53	2007	---	C	C	C	C	40 0	0	0 0	0 0 0	0 0	0111	-09	31	16	* 3170	01	00	062	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
54	2107	040	O	C	C	C	20 0	0	0 0	0 0 0	0 0	0101	* -00	23	06	2869	16	13	049	10	10	SC	040												10
55	2110	---	S	-B	C	C	15 0	0	0 0	0 0 0	0 0	0081	06	27	08	2864	27	22	041	6	2	AC	100	4	CI	300	6	0	0	---	6	* 0	0	--	3
56	2201	025	B	O	C	C	05 0	* 0	R- 0	0 0 8	F 0	9950	29	25	05	2829	29	29	100	10	6	SC	025	4	NS	040	10								10
57	2207	002	X	C	C	C	* 00 1/4	* 0	0 0	S 0 0	0 0	9924	20	29	20	2822	24	23	085	10	10	-	002												10
58	2216	---	C	C	C	C	30 0	0	0 0	0 0 0	0 0	9994	-04	29	31	2829 2850	14	11	044	0	0	SC	030	0	0	---	0	0	0	---	0	0	0	---	0
59	DUP1 2301	---	C	C	C	C	10 0	0	0 0	0 0 0	0 0	0110	-07	29	24	2871	07	05	052	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
60	DUP2 2301	---	C	C	C	C	15 0	0	0 0	0 0 0	0 0	0110 0269	00	28	08	2871 2917	09	08	067	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
61	2304	---	C	C	C	C	150 80 0	0	0 0	0 0 0	0 0	0269 0143	-07	28	14	2917 2880	06	04	055	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
62	2307	---	C	C	C	C	800 10 0	0	0 0	0 0 0	0 0	0174	-09	29	13	2889	04	02	054	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
63	MSG 2401																																		
64	2416	---	-O	C	C	C	15 0	0	0 0	0 0 0	0 0	0250	06	13	07	2912	005 26	21	042	10	10	CS	300	0	0	---	10	0	0	---	10	0	0	---	4

∗ COLUMN 3 ENTRIES: S = SCATTERED LAYER B = BROKEN LAYER X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS DENOTES OBSCURING PHENOMENA OTHER THAN FOG)
C = NO CLOUD O = OVERCAST LAYER

NOTE: "+" DENOTES "HEAVY"
"–" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
WBAN 10 REV. AUG. 1972

Fig. 3d

NOAA Form 82-384
(8-73)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

STATION NO.	YEAR	NO.	STATION NAME	PRESSURE TABLE USED
00001	70	08	WBAN EDIT TEST #1	29 . INCHES

	COLS.	2	3				4	5				6	8	9	10	17	18	19	20	21	22			25			28	29			32	33			36
			* SKY SYMBOLS					WEATHER AND/OR OBSTRUCTION TO VISION													CLOUDS AND OBSCURING PHENOMENA: LOWEST LAYER			2ND LAYER				3RD LAYER				4TH LAYER			
	DAY	CEILING (HDS. OF FT.)	S 1	S 2	S 3	S 4	VISIBILITY (MILES)	TORNADO & THUNDER	LIQUID PRECIP.	FROZEN PRECIP.	OBS. TO VISION	SEA LEVEL PRESSURE (MBS.)	DEW POINT (°F)	WIND DIR. TENS OF DEGREES	WIND SPEED (KNOTS)	STATION PRESSURE (INCHES)	DRY BULB (°F)	WET BULB (°F)	RELATIVE HUMIDITY (%)	TOTAL SKY COVER	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	2ND LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	3RD LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	OPAQUE
65	2501	005	0	C	C	C	*00 1/4	*0	0 0	S- 0 0	0 0	0196	19	07	11	2898	24	23	081	10	10	ST	005												10
66	2504	003	0	C	C	C	05 0	*0	R- 0	0 0 0	F 0	0158	22	08	12	2887	*24	24	092	10	10	ST	003												10
67	2510	002	0	C	C	C	01 0	0	0 0	S 0 0	F 0	0106 0027	30	04	10	2872 2849	32	31	092	10	10	ST	002												10
68	2513	001	X	C	C	C	00 1/4	0	0 0	S- 0 0	F 0	0027 9930	31	05	07	2849 2822	31	31	100	10	10	FOG	001												10
69	2516	000	X	C	C	C	00 0	0	0 0	0 SG 0	F 0	9930 9836	34	02	05	*2796	36	35	092	10	10	FOG	000												10
70	2519	070	S	0	C	C	08 0	*0	RW-0	0 SG-0	0 0	9821	40	19	12	*2793	*45	43	083	10	2	CF	010	10	AC	070	10	0	0	---	10	0	0	---	9
71	2522	038	0	C	C	C	10 0	0	0 0	0 0 0	0 0	9822	33	20	12	*2793	39	36	079	10	10	SC	038	0	0	---	10	0	0	---	10	0	0	---	9
72	2601	040	S	0	C	C	15 0	0	0 0	0 0 0	0 0	9835	28	23	17	*2796	36	33	073	10	3	SC	020	7	SC	040	10								10
73	2607	030	0	C	C	C	*00 1/2	*0	0 0	S- 0 0	0 0	9904	15	22	11	2816	27	24	061	10	10	SC	030												10
74	2616	030	B	B	C	C	*01 0	*0	0 0	0 SG+0	0 0	9993	06	30	18	2840	24	20	046	9	6	CU	030	3	CI	300	9	0	0	---	9	0	0	---	7
75	MSG 2704																																		
76	2716	----	B	C	C	C	35 0	0	0 0	0 0 0	0 0	0102 0199	08	28	11	2869	26	21	046	8	8	CI	300	0	0	---	8	0	0	---	8	0	0	---	4
77	2722	---	C	C	C	C	20 0	0	0 0	0 0 0	0 0	0223 0123	10	28	12	2873	23	20	057	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
78	DUP1 2804	042	0	C	C	C	15 0	0	0 0	0 0 0	0 0	0109	12	28	07	2872	25	22	058	10	10	SC	042												10
79	DUP2 2804	---	C	C	C	C	15 0	0	0 0	0 0 0	0 0	0112	012 -01	30	14	2866	025 12	10	056	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
80	2816	---	*				30 0	0	0 0	0 0 0	0 0	0119	13	33	17	2876	33	27	044	3	*3	CU	045	*0	0	---	*3	*0	0	---	3	*0	0	---	3

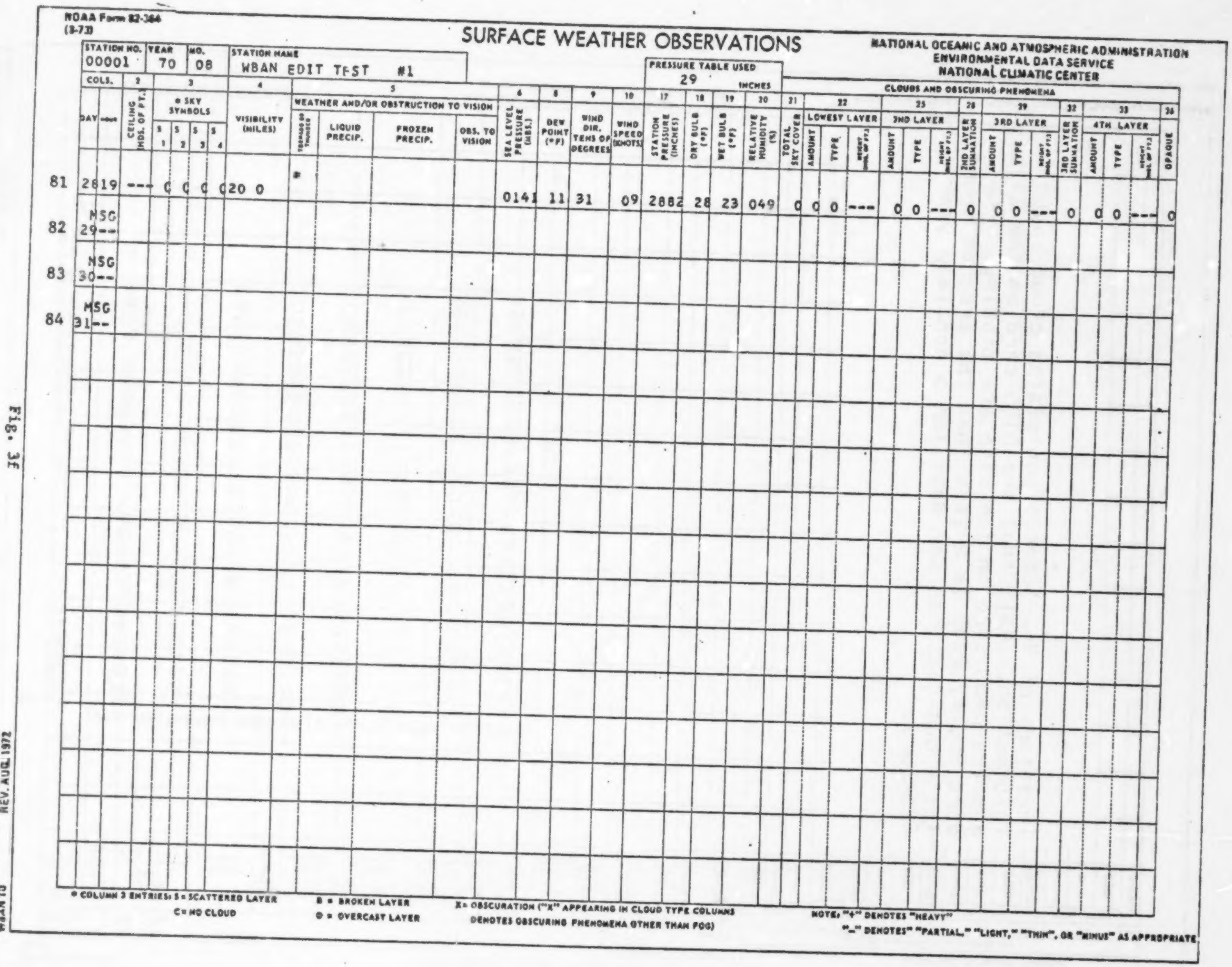
* COLUMN 3 ENTRIES: S = SCATTERED LAYER B = BROKEN LAYER
C = NO CLOUD O = OVERCAST LAYER

X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS DENOTES OBSCURING PHENOMENA OTHER THAN FOG)

NOTE: "+" DENOTES "HEAVY"
"–" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
REV. AUG. 1972
WBAN 10

Fig. 3e



NOAA Form 82-384
(3-73)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

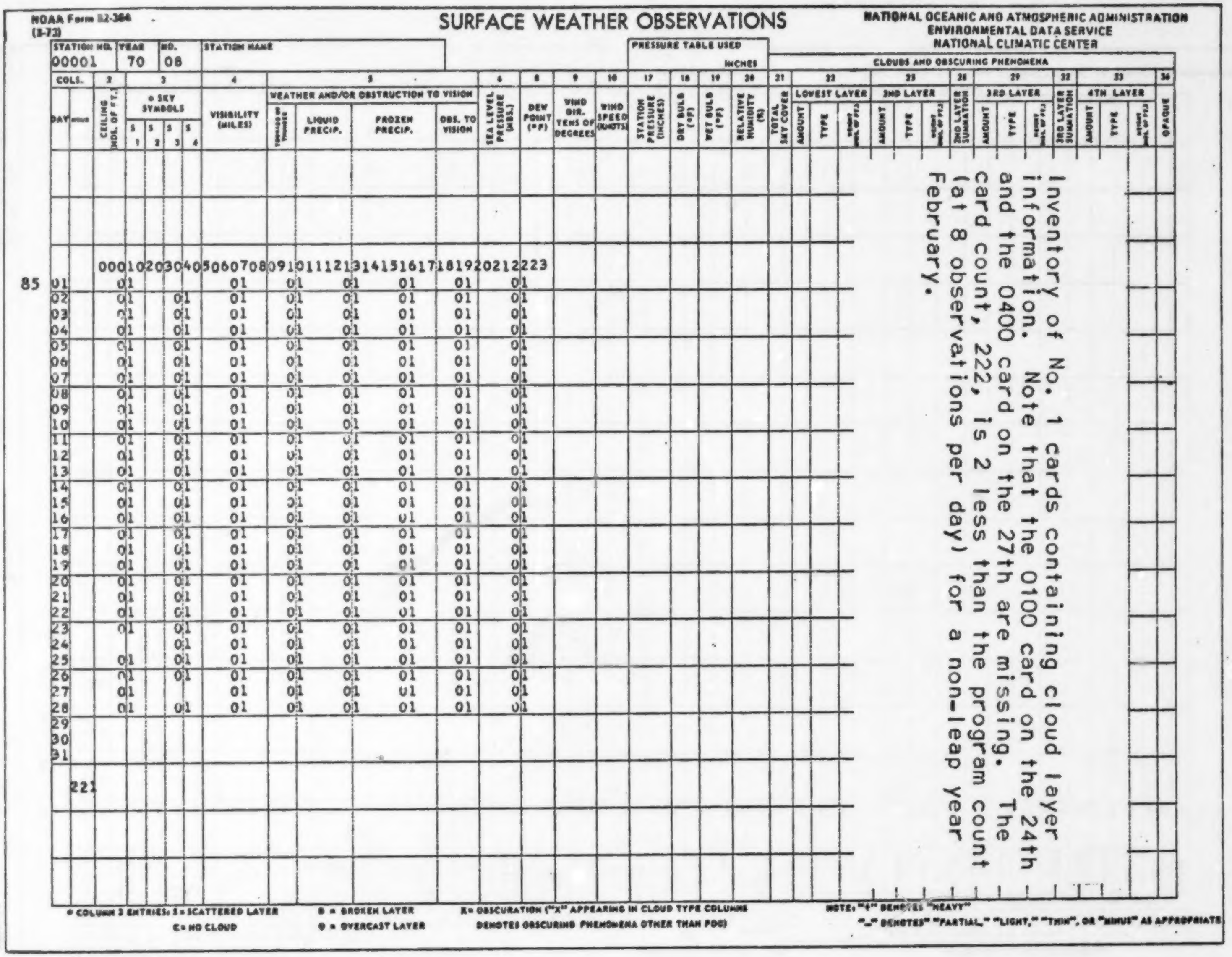
STATION NO.	YEAR	MO.	STATION NAME	PRESSURE TABLE USED
00001	70	08	WBAN EDIT TFST #1	29 INCHES

	COLS.		2	3				4	5				6	8	9	10	17	18	19	20	21	22			25			28	29			32	33			36
				* SKY SYMBOLS					WEATHER AND/OR OBSTRUCTION TO VISION													CLOUDS AND OBSCURING PHENOMENA														
																						LOWEST LAYER			2ND LAYER				3RD LAYER				4TH LAYER			
	DAY	HOUR	CEILING (HDS. OF FT.)	S 1	S 2	S 3	S 4	VISIBILITY (MILES)	TORNADO OR THUNDER	LIQUID PRECIP.	FROZEN PRECIP.	OBS. TO VISION	SEA LEVEL PRESSURE (MBS.)	DEW POINT (°F)	WIND DIR. TENS OF DEGREES	WIND SPEED (KNOTS)	STATION PRESSURE (INCHES)	DRY BULB (°F)	WET BULB (°F)	RELATIVE HUMIDITY (%)	TOTAL SKY COVER	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	2ND LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	3RD LAYER SUMMATION	AMOUNT	TYPE	HEIGHT (HDS. OF FT.)	OPAQUE
81	28	19	---	C	C	C	C	20 0	#				0141	11	31	09	2882	28	23	049	0	0	0	---	0	0	---	0	0	0	---	0	0	0	---	0
82	29	MSG --																																		
83	30	MSG --																																		
84	31	MSG --																																		

* COLUMN 3 ENTRIES: S = SCATTERED LAYER
C = NO CLOUD
B = BROKEN LAYER
O = OVERCAST LAYER
X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS DENOTES OBSCURING PHENOMENA OTHER THAN FOG)
NOTE: "+" DENOTES "HEAVY"
"–" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
WBAN 10 REV. AUG. 1972

Fig. 3f



NOAA Form 82-384
(3-72)

SURFACE WEATHER OBSERVATIONS

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ENVIRONMENTAL DATA SERVICE
NATIONAL CLIMATIC CENTER

STATION NO.	YEAR	MO.	STATION NAME
00001	70	08	

PRESSURE TABLE USED INCHES

COLS. 2 CEILING (HDS. OF FT.)
3 * SKY SYMBOLS S 1 S 2 S 3 S 4
4 VISIBILITY (MILES)
5 WEATHER AND/OR OBSTRUCTION TO VISION: THUNDER OR TORNADO, LIQUID PRECIP., FROZEN PRECIP., OBS. TO VISION
6 SEA LEVEL PRESSURE (MBS.)
8 DEW POINT (°F)
9 WIND DIR. TENS OF DEGREES
10 WIND SPEED (KNOTS)
17 STATION PRESSURE (INCHES)
18 DRY BULB (°F)
19 WET BULB (°F)
20 RELATIVE HUMIDITY (%)
21 TOTAL SKY COVER
CLOUDS AND OBSCURING PHENOMENA
22 LOWEST LAYER: AMOUNT, TYPE, HEIGHT (HDS. OF FT.)
25 2ND LAYER: AMOUNT, TYPE, HEIGHT
26 2ND LAYER SUMMATION
29 3RD LAYER: AMOUNT, TYPE, HEIGHT
32 3RD LAYER SUMMATION
33 4TH LAYER: AMOUNT, TYPE, HEIGHT
34 OPAQUE

DAY								
	000102030405060708091011121314151617181920212223							
01	01		01	01	01	01	01	01
02	01	01	01	01	01	01	01	01
03	01	01	01	01	01	01	01	01
04	01	01	01	01	01	01	01	01
05	01	01	01	01	01	01	01	01
06	01	01	01	01	01	01	01	01
07	01	01	01	01	01	01	01	01
08	01	01	01	01	01	01	01	01
09	01	01	01	01	01	01	01	01
10	01	01	01	01	01	01	01	01
11	01	01	01	01	01	01	01	01
12	01	01	01	01	01	01	01	01
13	01	01	01	01	01	01	01	01
14	01	01	01	01	01	01	01	01
15	01	01	01	01	01	01	01	01
16	01	01	01	01	01	01	01	01
17	01	01	01	01	01	01	01	01
18	01	01	01	01	01	01	01	01
19	01	01	01	01	01	01	01	01
20	01	01	01	01	01	01	01	01
21	01	01	01	01	01	01	01	01
22	01	01	01	01	01	01	01	01
23	01	01	01	01	01	01	01	01
24		01	01	01	01	01	01	01
25	01	01	01	01	01	01	01	01
26	01	01	01	01	01	01	01	01
27	01		01	01	01	01	01	01
28	01	01	01	01	01	01	01	01
29								
30								
31								
	221							

85

Inventory of No. 1 cards containing cloud layer information. Note that the 0100 card on the 24th and the 0400 card on the 27th are missing. The card count, 222, is 2 less than the program count (at 8 observations per day) for a non-leap year February.

* COLUMN 3 ENTRIES: S = SCATTERED LAYER B = BROKEN LAYER X = OBSCURATION ("X" APPEARING IN CLOUD TYPE COLUMNS DENOTES OBSCURING PHENOMENA OTHER THAN FOG)
C = NO CLOUD O = OVERCAST LAYER

NOTE: "+" DENOTES "HEAVY"
"–" DENOTES "PARTIAL," "LIGHT," "THIN", OR "MINUS" AS APPROPRIATE

FORM 5640 SURFACE OBSERVATIONS
REV. AUG. 1972
WBAN 10

Fig. 3g

V. Visual Checks

Data that are not keyed onto tape are given a limited visual scan as a random check for quality control and consistency of climatological data entries.

VI. WBAN No. 3 Card Edit and Listing

The No. 3 WBAN Card images (Fig. 4) contain varying daily Climatological data for individual stations for the period midnight-midnight LST.

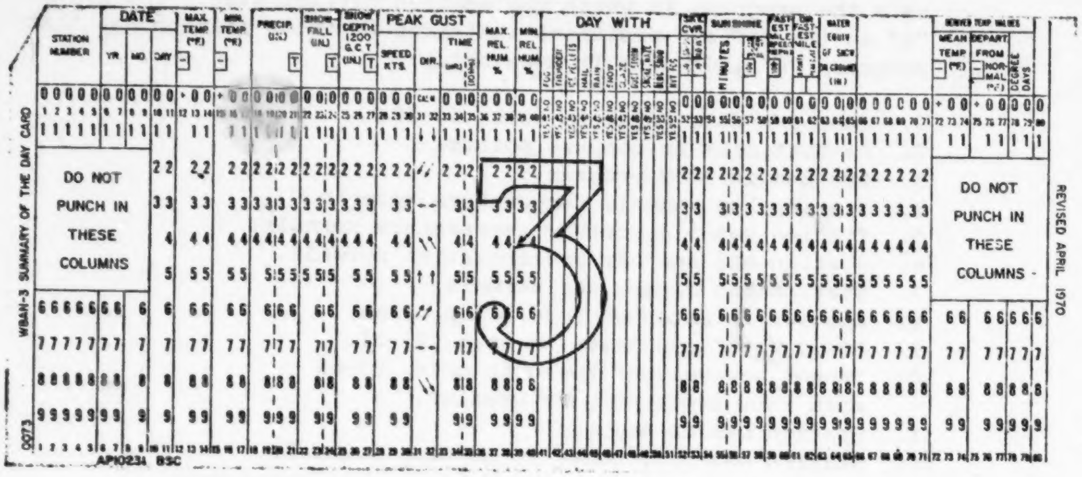

STATION NUMBER
DATE
YR
MO
DAY
MAX. TEMP (°F)
MIN. TEMP (°F)
PRECIP. (IN.)
SNOW-FALL (IN.)
SNOW DEPTH 1200 G.C.T. (IN.)
PEAK GUST
SPEED KTS.
DIR.
TIME
MAX. REL. HUM. %
MIN. REL. HUM. %
DAY WITH
SKY CVR.
SUNSHINE
MINUTES
FASTEST MILE
WATER EQUIV. OF SNOW ON GROUND (IN.)
MEAN TEMP (°F)
DEPART. FROM NORMAL (°F)
DEGREE DAYS
DO NOT PUNCH IN THESE COLUMNS
DO NOT PUNCH IN THESE COLUMNS
3
WBAN-3 SUMMARY OF THE DAY CARD
REVISED APRIL 1970
9073
AP1023 BSC

Fig. 4

A. General

The edit of taped data is for internal consistency and corrected, if necessary. The corrected (updated) tape is used in preparing listings for verification of Form F-6 (for stations submitting them) and in the preparation of formats, tapes, and summary cards used in the LCD and CD programs and for archiving of corrected data for future use.

VII. Program and Tape Control Procedures

In order to provide appropriate edit information and to meet publication deadlines, the records are placed on two separate tapes each month, preceded by certain station program and priority information.

A. Tape No.1 contains data for all stations in the LCD and CD programs and Tape No.2 contains data for all others.

B. A thirteen-digit Program and Priority Editing Code is provided in the station's header on the name tape. Positions 1-11 indicate the programs in which the station participates. The figure "1" in the various positions indicates that the station has that program; "0" indicates it does not.

1. Sunshine data keyed in cols. 54-58.
2. Fastest mile in compass points.
3. Station in 1009 program.
4. Station has monthly temperature normals.
5. Station has mid-monthly temperature normals.
6. Station has monthly precipitation normals.
7. Station has degree day normals.
8. Station in Extended Forecast program.
9. "Days With" are keyed in cols. 41-51.
10. "Water Equivalent" keyed in cols. 63-65 when snow depth 002, or greater.
11. Station has CD number.
12. Station has LCD, coded 1; no LCD, coded 2.
13. Station is operating, coded 1; closed, coded zero (a convenience in using the name tape as a reference in other programs).

VIII. Edit and Listing

The machine edit is designed to detect various inconsistencies of data. The corrected (updated) listings provide various computations of sums, averages, departures from normal and counts of number of occurrences, etc., used in the climatological programs.

Sample listings appear on pages 20a and 20b.

The fields for all inconsistencies noted in the edit are flagged with appropriate symbols in the column(s) to the right of the field(s) questioned.

Checks are made and field(s) flagged for review, according to the outline below:

A. All columns 1-80

"12" overpunch.

WB-	YR MO	STATION NAME	MACHINE CONTROL CODES	STATE INDEX DIV
00001	67 02	TEST STATION	0010110110111	0

DAY	MAX TEMP	MIN TEMP	MEAN TEMP	DEGREE DAYS	PCPN	SNOW: FALL	ON GRND	WATER EQUIV (1)	FAST SPD	MILE DIR (2)	SUNSHINE HOURS	%	SKY S-S (3)	COVER M-M	F	T	IP	A	R	S	Z	D	KH	BS	HF	PEAK GUST SPD	DIR	TIME
01	44	38	41	24	.00	2.1*	2*			036 0	11.0	0	1	2	0	0	1*	1*	1*	1*	1*	0	0	0	1*	100		0030
02	37	38*	38	27	M	.0	1		100	0	9.8	1000	2	6*	0	0	1*	0	0	1*	1*	0	0	0	0		011	0001
03	46	45*	46	19	1.20	20.1*	1*	.5	51*	2	8.7	0	1	3	0	0	1*	0	1	1*	1*	0	0	0	0	51*	NE	043
04	- 1*	- 9	- 5	70	.15	1.0*	0		22	NNE0+	9.0	0	3	4	0	0	0	0	1	0	0	0	0	0	0		*	*166
05	140	- 1	70		.08	.0	1		19	31 *	4.0	1000	6	2	0	0	0	0	1	0	0	0	0	0	0	19	W	154
06	60	3 0			.060	.60	60	.0	0102*	37 *	1 .10	0	20	30	20	30	40	5*	6*	70	80	9*	10	20	20	21*37	0	430
07M																												
08	51	39	45	20	.92	.0	3*	.4	23	33	2.6	25	9	9	1	0	0	0	1	0	0	0	0	0	1	28	NW	201
09	44	25	35	30	.00	.0	1		21	33	10.3	100	1	3	0	0	0	0	0	0	0	0	0	0	0	22	NW	007
10	29	23	26	39	.23	.4	1		15	8	.0	0	10	10	1	0	1	0	1	1	1	0	0	0	1	16	E	084
11	46	27	37	28	T	.0	1		18	27	7.8	75	8	8	1	0	0	0	1	0	1	0	0	0	1	22	W	063
12	44	31	38	27	.08	T	1		13	23	.0	0	10	10	1	0	1	0	1	0	1	0	0	0	1	16	SW	239
13	45	19	32	33	.00	.0	1		29	29	8.2	78	2	3	1	0	0	0	0	0	0	0	0	0	1	31	WNW	059
14	20	12	16	49	T	T	1		13	29	.0	0	10	9	0	0	0	0	0	1	0	0	0	0	0	15	WNW	009
15	34	12	23	42	.03	.5	1		12	26	9.9	93	0	2	0	0	0	0	0	1	0	0	0	0	0	15	SW	220
16	44	20	32	33	.00	.0	1		21	29	10.2	95	6	7	0	0	0	0	0	0	0	0	0	0	0	28	WNW	203
17	36	17	27	38	.00	.0	1		12	22	10.7	100	0	3	0	0	0	0	0	0	0	0	0	0	0	16	SSW	227
18	37	22	30	35	.01	.1	1		22	29	1.6	14	9	8	1	0	0	0	0	1	0	0	0	0	0	27	WNW	162
19	22	2	12	53	.00	.0	1		38	32	8.6	80	6	4	0	0	0	0	0	0	0	0	0	0	0	43	NW	201
20	21	1	11	54	.00	.0	1		27	31	10.8	100	0	1	0	0	0	0	0	0	0	0	0	0	0	38	NW	006
21	45	13	29	36	.07	T	1		17	29	6.8	63	5	7	0	0	0	0	1	1	0	0	0	0	0	19	WNW	006
22	37	7	22	43	.25	2.6	1		36	29	5.4	50	5	5	1	0	0	0	1	1	0	0	0	1	0	43	W	162
23	20	4	12	53	.00	.0	3	.3	28	29	10.8	98	0	0	0	0	0	0	0	0	0	0	0	0	0	32	W	042
24	28	8	18	47	.00	.0	3	.3	12	6	10.0	91	6	5	0	0	0	0	0	0	0	0	0	0	0	14	SSE	181
25	50	23	37	28	2.04	T	2	.2	22	5	.0	0	10	-10	1	1	1	0	1	1	1	0	0	0	1	31	S	168
26	37	16	27	38	T	T	1		24	27	5.1	46	8	7	0	0	0	0	0	1	0	0	0	0	0	31	W	018
27	28	10	19	46	.00	.0	1		23	30	10.8	97	3	3	0	0	0	0	0	0	0	0	0	0	0	29	NW	087
28	36	22	29	36	.00	.0	1		23	30	6.5	58	4	4	0	0	0	0	0	0	0	0	0	0	0	26	NNW	164
29M																												
30M																												
TOTAL	1020	464		948	5.06*	26.8					178.6		125	135	8	1	6	1	11	11	7			1	7			
AVERAGE	39.2	17.8	28.5										4.8	5.2														
DEP. FROM NORMAL			3.6	-175	2.14																							
EXTREME DATE	140 5	- 9 4			2.04 25	20.1 3	3 24+		51	2 3																51 3	NE	

NUMBER OF DAYS:

Max. ≤ 32°	Max. ≥ 90°$	Min. ≤ 32°	Min. ≤ 0°
8	1	22	2

$ ≥ 70° at Alaskan stations.

PRECIPITATION T	.01+	.10+	.50+	1.00+
3	11	6	3	2

+ Indicates measurement ≥ threshold value.

SNOW 1.0+
4

CLEAR	PTLY CLDY	CLDY
11	7	8

(1) Water equivalent of snow and ice on ground.

(2) If the / (solidus) appears, speeds are gusts. Figures for directions are tens of degrees from true North; i.e., 9 = East, 18 = South, 27 = West, and 36 = North. When directions are in tens of degrees, speeds are fastest observed 1-minute values.

(3) S-S indicates sunrise to sunset and M-M midnight to midnight.

(4) Entry of 1 indicates occurrence, 0 indicates no occurrence. Weather types are: F = fog, visibility more than ¼ mile; T = thunderstorm; IP Ice P.; A = hail; R = rain; S = snow; Z = glaze; D = dust, visibility ½ mile or less; KH = smoke or haze or both; BS = blowing snow; and HF = heavy fog (visibility ¼ mile or less due to fog).

Reproduced from best available copy.

.WBAN # 14836 YR MO 75 06 STATION NAME LANSING MICHIGAN WBAS MACHINE CONTROL CODES 11111111111111 STATE INDEX DIV 20 4641 9

FIELD: 1 2 3 4 5 6 7 8 9 10 11 12 13

DAY	MAX TEMP	MIN TEMP	MEAN TEMP	DEGREE DAYS*	PCPN	SNOW FALL	ON GRND	WATER EQUIV (1)	FAST SPD	MILE DIR (2)	SUNSHINE HOURS	SUNSHINE %	SKY COVER S-S (3)	SKY COVER M-M	F	T	IP	A	R	S (4)	Z	D	KH	BS	HF	PEAK SPD	GUST DIR	GUST TIME	COOL DEG
01	72	45	59	6	.25	.0	0		16	SW	7.4	49	7	5	0	1	0	0	1	0	0	0	0	0	0	20	SW	140	
02	70	42	56	9	.00	.0	0		11	S	9.5	63	7	5	0	0	0	0	0	0	0	0	0	0	0	17	SW	169	
03	72	42	57	8	.00	.0	0		18	W	10.2	67	7	5	0	0	0	0	0	0	0	0	0	0	0	22	W	174	
04	74	50	62	3	.24	.0	0		22	SE	2.4	16	9	7	1	1	0	0	1	0	0	0	0	0	0	25	S	165	
05	76	59	68		.38	.0	0		30	W	6.2	41	7	7	1	1	0	0	1	0	0	0	0	0	0	32	W	017	3
06	68	54	61	4	.00	.0	0		27	NW	9.2	61	5	5	0	0	0	0	0	0	0	0	0	0	0	29	W	101	
07	60	51	56	9	.00	.0	0		19	NW	.1	1	10	9	0	0	0	0	0	0	0	0	0	0	0	21	NW	026	
08	66	44	55	10	.00	.0	0		11	NE	10.5	69	4	3	0	0	0	0	0	0	0	0	0	0	0	15	NE	133	
09	76	40	58	7	.00	.0	0		15	SE	13.3	87	6	4	1	0	0	0	0	0	0	0	0	0	0	21	SE	147	
10	78	52	65		.00	.0	0		24	SE	13.7	90	5	5	0	0	0	0	0	0	0	0	0	0	0	24	E	114	0
11	70	58	64	1	.50	.0	0		17	SW	.0	0	10	9	1	0	0	0	1	0	0	0	1	0	0	21	SW	227	
12	76	60	68		.02	.0	0		27	W	8.5	56	5	7	0	0	0	0	1	0	0	0	1	0	0	32	W	194	3
13	80	60	70		T	.0	0		21	W	11.5	75	3	6	0	0	0	0	1	0	0	0	0	0	0	26	SW	225	5
14	81	59	70		.25	.0	0		21	SW	8.8	57	7	7	0	1	0	0	1	0	0	0	0	0	0	22	SW	004	5
15	76	60	68		.50	.0	0		54*	NW	1.6	11	10	10	1	1	0	0	1	0	0	0	0	0	0	60*	SW	126	3
16	75	56	66		T	.0	0		17	[illegible]	9.0	59	6	7	0	0	0	0	1	0	0	0	0	0	0	22	W	014	1
17	83	66	75		.11	.0	0		28	SW	4.2	27	9	9	0	0	0	0	1	0	0	0	0	0	0	33	SW	172	10
18	85	68	77		T	.0	0		15	SW	10.2	67	6	6	1	1	0	0	1	0	0	0	1	0	0	18	SW	013	12
19	88	71	80		.01	.0	0		16	SW	7.1	47	7	7	1	1	0	0	1	0	0	0	1	0	0	20	SW	004	15
20	85	68	77		T	.0	0		13	E	7.0	45	8	7	0	1	0	0	1	0	0	0	1	0	0	15	E	147	12
21	88	67	78		.00	.0	0		13	SW	10.7	70	3	4	0	0	0	0	0	0	0	0	1	0	0	14	SW	163	13
22	92	69	81		.00	.0	0		18	SW	10.0	65	6	5	0	0	0	0	0	0	0	0	1	0	0	20	SW	191	16
23	86	73	80		.00	.0	0		20	W	10.5	68	6	6	0	0	0	0	0	0	0	0	1	0	0	22	W	110	15
24	78	65	72		.67	.0	0		17	NE	1.7	11	10	10	0	1	0	0	1	0	0	0	1	0	0	24	NE	128	7
25	77	61	69		.00	.0	0		11	E	4.8	32	9	9	1	0	0	0	0	0	0	0	0	0	0	14	NE	163	4
26	82	57	70		.00	.0	0		8	E	14.5	95	1	1	0	0	0	0	0	0	0	0	1	0	0	13	NE	129	5
27	86	62	74		.00	.0	0		7	E	7.1	47	8	6	1	0	0	0	0	0	0	0	1	0	0	11	E	116	9
28	88	67	78		.00	.0	0		10	E	6.8	44	6	5	1	0	0	0	0	0	0	0	1	0	1	13	E	178	13
29	87	60	74		.00	.0	0		12	E	13.2	86	2	2	1	-0	0	0	0	0	0	0	1	0	0	16	NE	159	9
30	86	53	70		.00	.0	0		10	NE	15.3	100	2	1	0	0	0	0	0	0	0	0	0	0	0	12	NE	131	5
TOTAL	2361	1739		H 57 S 6970	2.93	.0					58.2 245.0		191	179	11	9	0	0	14	0	0	0	13	0	1				M 165 S 238
AVERAGE	78.7	58.0	68.4								53	53	6.4	6.0															
DEP. FROM NORMAL			1.35	H 9 66	-.54																								M 54 S 101
EXTREME DATE	92 22	40 9			.67 24	.0	0		54	NW 15																60	SW 15		

NUMBER OF DAYS:

Max ≥ 90° †	Max ≤ 32°	Min ≤ 32°	Min ≤ 0°
1	0	0	0

† ≥ 70° at Alaskan stations.

PRECIPITATION T	.01+	.10+	.50+	1.00+	SNOW 1.0+	CLEAR	PTLY CLDY	CLDY
4	10	8	3	0	0	5	16	9

+ indicates measurement ≥ threshold value.

* BASE 65°F. ENTRY PRECEDED BY "C" INDICATES COOLING DEGREE DAYS, OTHERWISE HEATING DEGREE DAYS.

(1) Water equivalent of snow and ice on ground.
(2) If the / (solidus) appears, speeds are gusts. Figures for directions are tens of degrees from true North; i.e., 9 = East, 18 = South, 27 = West, and 36 = North. When directions are in tens of degrees, speeds are fastest observed 1-minute values.
(3) S-S indicates sunrise to sunset and M-M, midnight to midnight.
(4) Entry of 1 indicates occurrence, 0 indicates no occurrence. Weather types are: F = fog, visibility more than 1/4 mile; T = thunderstorm; IP = ice pellets; A = hail; R = rain; S = snow; Z = glaze; D = dust, visibility 1/2 mile or less; KH = smoke or haze or both; BS = blowing snow; and HF = heavy fog (visibility 1/4 mile or less due to fog).

National Oceanic and Atmospheric Administration EDS-NCC

NOAA Form 82-32?

Summary of the Day
Form No. 5670 Rev. 12-71

B. Day (col. 10-11)

   1. > Possible for month
   2. Missing

C. Max. Temp. (cols. 12-14)

   Legal punches are: X, 0, or 1 in col. 12 and 0-9 in cols. 13-14.

   1. Illegal punches
   2. < Min. Temp (cols. 15-17)
   3. < Min. Temp. (cols. 15-17) of previous day
      Print negative values (X punch in col. 12) with a minus (-) preceding numerical values in cols. 13-14.

D. Min. Temp. (cols. 15-17)

   Legal punches are: 0 or X for col. 15 and 0-9 for cols. 16-17.

   1. Illegal punches
   2. > Max. Temp. (cols. 12-14) of previous day
      Print negative values (X punch in col. 15) with a minus (-) preceding numerical values in cols. 16-17.

E. Precipitation (cols. 18-21)

   Legal punches are 0-9 or BBBX

   1. Illegal punches
   2. "0000" with cols. 22-24 other than "000" or "BBB"
   3. Col. 21 "X" with cols. 18-20 other than B
   4. Any of cols. 18-20B with col. 21 "0-9"
   5. > 1000
      Print BBBX as "T." Also print "000X" as "T" but flag as error as indicated above.

F. Snowfall (cols. 22-24)

   Legal punches are 0-9 or BBX

   1. Illegal punches
   2. Col. 24 "X" with cols. 22 and 23 other than B
   3. Cols. 22 or 23 B with col. 24 "0-9"
   4. > 200
      Print BBX as "T." Also print "00X" as "T," but flag as error as indicated above.

G. Snow Depth (cols. 25-27)

   Legal punches are 0-9 or BBX. May also be B for entire field.

   1. Illegal punches
   2. Other than "000" with cols. 22-27 for preceding day and cols. 22-24 for same day punched all 0's.

3. Col. 27 "X" with cols. 25 and 26 other than "B"

4. Cols. 25 or 26 B with col. 27 "0-9"
   Print "BBX" as "T." Also print "OOX" as "T," but flag as error as indicated above.

H. Peak Gusts, Direction and Time (cols. 28-35)

Legal punches are:

0 - 9 for cols. 28-30,
The "Alpha" Compass Point Code for cols. 31-32, and
000 - 239 for cols. 33-35 or entire field may be "B."
An "X" in col. 31 is programmed to convert peak gust speeds from knots to mph and publish under fastest mile heading with "/" following the direction as an indicator of peak gust speed. Omission of "X" in col. 31 will be flagged by "$" following the direction on the edit listing.
A "#" following the speed and direction spaces on the edit calls attention to entry of speed with direction omitted.

The Alpha Compass Point punches are:

00	C (calm)	22	NE	44	SE	66	SW
11	N	32	ENE	54	SSE	76	WSW
12	NNE	33	E	55	S	77	W
18	NNW	34	ESE	56	SSW	78	WNW

1. Illegal punches
2. Cols. 28-30 > 050
   Print in Alpha Code Letters.

I. "Days With" (cols. 41-51)

Legal punches are 0 or 1 if in station's program, otherwise all cols. should be B. (If punched, all columns should be punched.)

1. Illegal punches
2. Col. 41 "0" with "1" in col. 51
3. Col. 43 "1" with either or both fields (cols. 18-21, 22-24) all 0's.
4. Col. 43 "1" with min. temp. (cols. 15-17) > 044
5. Col. 44 "1" with cols. 18-21 "0000"
6. Col. 44 "1" with cols. 43 & 46 "0" & cols. 22-24 other than "0000"
7. Col. 45 "1" with cols. 18-21 "0000"
8. Col. 46 "1" with either or both fields (cols. 18-21, 22-24) all 0's.
9. Col. 46 "1" with min. temp. (cols. 15-17) > 044
10. Col. 47 "1" with col. 45 "0" (some exceptions, but flag)
11. Col. 47 "1" with min. temp. (cols. 15-17) > 039
12. Col. 50 "1" with either cols. 28-30 or 59-60 (if punched 010 or 10 respectively).

J. Sky Cover (cols. 52, 53)

Legal punches are 0-9 and X for both cols. or "B" for col. 53 if cols. 41-51 are B.

1. Col. 52 B with other than B in col. 53
2. Col. 52 > 3 greater than col. 53
3. Col. 53 other than B with cols. 41-51 B
4. Col. 53 > 2 greater than col. 52

Print "X" punches as "10"

K. Sunshine and Percent of Possible (cols. 54-58)

Legal punches are: 000-199 for cols. 54-56, 0-9 or X for col. 57 and 0-9 or B for col. 58. Also entire field may be B.

1. Illegal punches
2. Col. 57 "X" with underpunch
3. Cols. 54-58 are blank
4. Col. 57 "X" with other than B in col. 58

Print as "100" when cols. 57-58 punched "XB." Also print "100 when col. 57 has an "X" punch regardless of other illegal punching in either or both cols. 57 or 58, but flag as error as indicated above.

5. With cols. 54-56 punched 000, cols. 57-58 with other than zeros
6. With cols. 57-58 punched 000, cols. 54-56 with other than zeros
7. With cols. 54-56 punched greater than 000, cols. 57-58 will be greater than 00.

L. Fastest Mile and Direction (cols. 59-62)

Legal punches are: 0-9 for cols. 59-60 with an X overpunch permitted in col. 59 for speeds of ≥ 100, 00-36 for cols. 61-62 if neither col. has an "X" overpunch, and the "Alpha" Compass Point Code (see VIII,H above) if either or both (cols. 61-62) have an "X" overpunch.

Illegal punches:

1. Cols. 59-60 "00" (without "X" overpunch in col. 59) with other than "00" in cols. 61-62.
2. Cols. 59-60 > 50.
3. Cols. 59, 61 or 62 punched "X" without an underpunch 0-9.
4. Col. 62 "X" overpunched with no "X" overpunch in col. 61.

Print:

1. "1" preceding speed punched in cols. 59-60 when col. 59 has an "X" overpunch.

2. Direction in the "Alpha" code letters when either or both cols. 61, 62 have an "X" overpunch. (See VIII, H above.)

3. A dash (-) in col. following direction with col. 61 has an "X" overpunch.

4. A plus (+) in 2nd col. following the direction when col. 62 has an "X" overpunch.

M. Water Equivalent (cols. 63-65)

Legal punches are: 0-9 or B. Water Equivalent is in inches & tenths.

Illegal punches: B in any of cols. 63-64 with col. 65 punched 0-9.

Other cols. (36-40, 66, 68-80) should be B.

IX. Machine Computations

Various sums, means, departures (from pre-programmed normals), frequency counts, summary cards, etc., necessary in the verification program and used in the preparation of formats for the LCD, CDNS, and Table J are made by the computer.

Print the sums, averages, etc., from the data available when some days and/or items are missing.

A. Daily Computation are made for:

1. Average temperature
2. Departure from normal
3. Degree days

B. Monthly Sums are computed and listed for:

1. Max. temperature
2. Min. temperature
3. Mean temperature
4. Degree days, heating and cooling
5. Precipitation
6. Snowfall
7. Sunshine
8. "Days With" (if in station's program)
9. Sky Cover (SR-SS & Mid-Mid)

C. Monthly Averages are computed and listed for:

1. Max. temperature
2. Min. temperature
3. Mean temperature (this is 1/2 the sum of the average max. and min., C, 1 & 2 above).
4. Average percent of possible sunshine (sum of daily percentages divided by the number of days).

Monthly percent of possible sunshine is computed from total sunshine recorded and the pre-programmed possible amount, sunrise to sunset.

D. Monthly Departures are computed and listed for:

1. Mean Temperature
2. Degree days, heating and cooling
3. Precipitation

E. Seasonal Departure for Degree Days (from seasonal totals carried forward from preceding month and current month's total) are computed and listed. Season begins with July for heating and January for cooling.

F. Extremes and Dates are selected and listed for:

1. Highest temperature
2. Lowest temperature
3. Greatest precipitation
4. Greatest Snowfall
5. Greatest Snow Depth
6. Greatest Wind Speed and Direction

(When the same value occurs on two or more dates, the date of the last occurrence followed by a plus (+) is listed. Also, the direction of the last occurrence of multiple "Greatest Wind speed" is printed.)

G. Frequency Counts are made and listed for:

1. Temperature

    a. Max. ≤ 32
    b. Max. ≥ 90, except ≥ 70 for Alaskan stations
    c. Min. ≤ 32
    d. Min. ≤ 00

2. Precipitation

    a. Trace (BBBX)
    b. $\geq$ 0001
    c. $\geq$ 0010
    d. $\geq$ 0050
    e. $\geq$ 0100

3. Snowfall

    a. $\geq$ 010

4. Character of Day (SR-SS)

    a. Clear (Avg. 0-3)
    b. Partly Cloudy (Avg. 4-7)
    c. Cloudy (Avg. 8-10) (Punched 8, 9, or X)

X. Precipitation Data Card Images

A. Program Involved. Hourly precipitation, monthly extremes, and maximum precipitation.

1. Hourly precipitation, greatest amounts of precipitation, snowfall, and snow depth and maximum precipitation are contained in a series of tape formats currently known as the HPD Deck. These are identified as to station, year, and month in the same manner as the WBAN #1 and #3 cards.

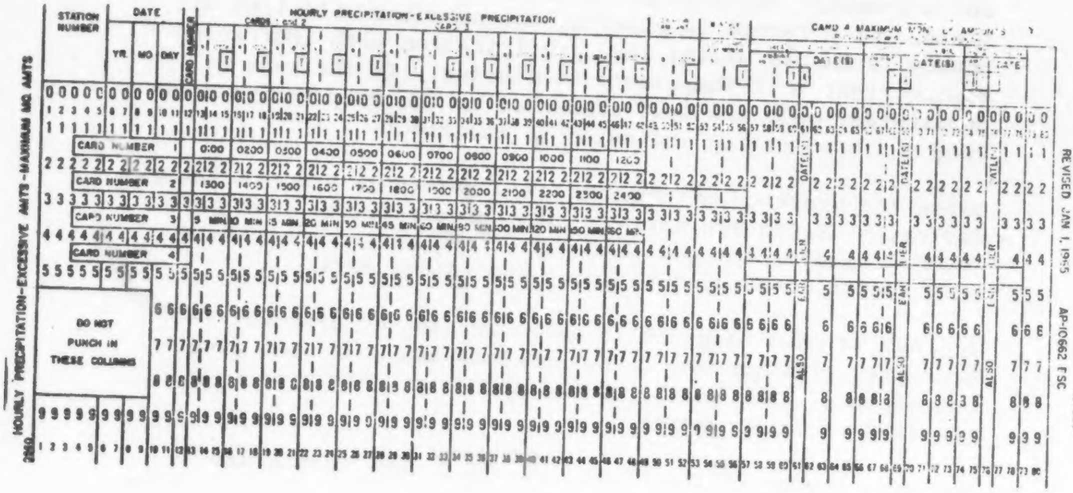

HOURLY PRECIPITATION-EXCESSIVE AMTS-MAXIMUM MO AMTS
2860
STATION NUMBER
DATE
YR
MO
DAY
CARD NUMBER
HOURLY PRECIPITATION-EXCESSIVE PRECIPITATION
CARD NUMBER 1
0100 0200 0300 0400 0500 0600 0700 0800 0900 1000 1100 1200
CARD NUMBER 2
1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400
CARD NUMBER 3
5 MIN 10 MIN 15 MIN 20 MIN 30 MIN 45 MIN 60 MIN 80 MIN 100 MIN 120 MIN 150 MIN 180 MIN
CARD NUMBER 4
CARD 4 MAXIMUM MONTHLY AMOUNTS
DATE(S)
DATE(S)
DATE
ALSO
ALSO
ALSO
DO NOT PUNCH IN THESE COLUMNS
REVISED JAN 1, 1965
AP-10662 FSC

Fig. 5

2. Hourly precipitation, HPD card format 1 or 2 in col. 12 as identifier.

For each station in the LCD program, #1 and 2 HPD data are keyed each day with precipitation and for the last day of the month whether precipitation has occurred or not. If stations are not equipped with recording gages, amounts are keyed only at 6-hourly synoptic times. In this case, the daily total is not keyed in the second format; the monthly total, however, is keyed in the last format of the month for all stations.

B. Checking Procedure - Hourly Precipitation

1. The checking is accomplished by a computer cross-foot listing to insure internal compatibility. A second check is made between the daily totals and the monthly total. The cross-foot for each station is begun by building in the memory of the computer a grid of zeros for all days in the month, i. e., 28 days, 30 days, etc., as the calendar requires. The keyed data are read into the grid and then edited. Information concerning missing record, blank fields, erroneous keying, and arithmetic mistakes is listed to the right of the data field. If a record is missing, the computer will list all hourly fields as having zero precipitation with indication to the right that the record is missing. In the case of duplicates, only the last presented to the computer will be used and duplication indicated to the right.

   Since the presence of the HPD 1 & 2 record is a controlling factor, stations not having hourly precipitation must have "dummy" records for the last day of the month, containing only identification, date, and card number data.

   The HPD #4 card image (4 in col. 12) has the greatest 24 hour precipitation and date, snowfall and date, and greatest depth of snow on the ground and date. There is only one #4 per station month.

2. The edit checks of the HPD 1, 2, and 4 card images are as follows: (Sample shown on page 179.

Column	Data	Edit Check
1- 5	Station No.	Sequence checked by number with a 4 punched in column 12 of 1st image.
6- 9	Year & Month	Values are checked and must be the same for the entire edit. Month must be in range of 01-12 in cols. 8-9.
10-12	Day Card No.	Only days with pcpn. are keyed except for the last day of the month. Each day will have only two images identified as 1 and 2 in col. 12. No. 2 has the daily total in cols. 49-52. The #4 in col. 12 will not have day punched in cols. 10-12.

Hourly edit

```
                 1   2   3   4   5   6   7   8   9   10  11  12
00001 70 06  1 1 000 000 005 00- 000 000 001 000 000 000 000 000
               2 000 000 000 000 004 004 000 000 000 000 000 000  0010  ERROR HR  3 4
00001 70 06  2 1 000 000 000 000 000 000 000 000 000 000 000 000        MISSING NO. 1 CARD
               2 000 000 003     000 000 000 000 000 000 000 000  0003  ERROR HR     16
00001 70 06  4 1 000 000 000 000 000 000 000 000 000 000 000 000        MISSING NO. 1 CARD
               2 000 000 000   T 000 000 000 000 000 000 000 000  00 -  CROSSFOOT ERROR    T
00001 70 06  5 1 000 000 000 000 000 0 1 000 000 000 001 000 000
               2 000 000 000 000 000 009 000 000 000 000 000 000  0006  ERROR HR         6
00001 70 06  6 1 000 000 000 000 000 000 000 000 000 000 000 000        MISSING NO. 1 CARD
               2 000 000 000 000  -  000 000 000 000 000 000 000     T  ERROR HR          17
00001 70 06  7 1 000 000 001 005 005 005 005 005 005 000 000 --5
               2 000 000 005 000 005 005 000 000 000 000 000 000        ERROR HR           12
00001 70 06  8 1 002 010 005 002 003 000 000 005 005 00  000 000
               2 000 000 010 015 005  10 010 005 005 000 000 000  0092  ERROR HR                1018
00001 70 06 10 1 000 000 003 005 005 000 000 000 005   T 005 000
               2 000 000 000 000 000 000 000 000 000 000 000 000  0023
00001 70 06 11 1 000 000 000 000 000   T   T 000 000 000 000 000
               2 000 000 000 000 000   T   T 000 000 000 000 000     T
00001 70 06 12 1 000 000 001 001 001 001 000 000 000 000 000 000
               2 000 000 001 001 001 001 000 000 000 000 000 000  0008
00001 70 06 14 1 000 000 000   T 000 000 000 000   T 000 000 000
               2 000 000 000   T 000 000 000 000   T 000 000 000     T
00001 70 06 15 1 000 000 000 000 001 000 000 000 000 000 000 000
               2 000 000 000 000 001 000 000 000 001 000 000 000  0003
00001 70 06 18 1 000 000 000 000   T   T 000 000 000 000 000 000
               2 000 000 000 000   T   T 001 000 000 000 000 000  0001
00001 70 06 21 1 000 000 001 001 001 000 000 000 000 000 000 000
               2 000 000 001 001 001 000 000 001 000 000 000 000  0007
00001 70 06 22 1 000 000 001 001 001 000 000 001 001 000 000 000
               2 000 000 005 005   T 010 000 000 000 000 000 000  0025
00001 70 06 25 1 004 010 010 015 005 020 015 020 010 005 005 010
               2 005 010 010 010   T 005 005 010 005 005 000 010  0204
00001 70 06 26 1 000 000   T 000 000 000 000 000   T 000 000 000
               2 000 000   T 000 000 000 000 000   T 000 000 000     T
00001 70 06 30 1
               2
00001 70 06    4                                                   0204 00025 026 00022 008 -07
CARD MONTHLY TOTAL 0430  COMPUTED TOTAL 0380  ERROR
COMPUTED HIGH-24 PRECIP 0204  DATES  25
```

6 Hourly Edit

```
                 1   2   3   4   5   6   7   8   9   10  11  12
23237 70 06  8 1             000             000
               2             001               T                  CROSSFOOT ERROR 0001
23237 70 06  9 1             005             001
               2             000             000                  CROSSFOOT ERROR 0006
23237 70 06 12 1             000             000
               2             000               T                  CROSSFOOT ERROR    T
23237 70 06 14 1               T               T
               2             000             000                  CROSSFOOT ERROR    T
23237 70 06 25 1             000             000
               2             000               T                  CROSSFOOT ERROR    T
23237 70 06 26 1             000               T
               2             000             000                  CROSSFOOT ERROR    T
23237 70 06 30 1
               2
23237 70 06    4                                                  0007 00509 000      000
CARD MONTHLY TOTAL 0007  COMPUTED TOTAL 0000  ERROR
NO HIGH-24 PRECIP COMPUTED
```

HPD EDIT LISTING

Column	Data	Edit Check
13-48	Hourly Values	Each hour has three cols. for data, i. e., hour 0100 cols. 13-15, etc. Zero pcpn. is keyed "000." BBO, BOO, OBO, OOB & BBB are flagged. All cols. are keyed with zeros placed to fill the col. Blanks are flagged. Trace amounts are indicated by an X in the right col. of the hour, preceded by two blank columns. Punched data of 00X, OBX, BOX and over-punches are flagged as errors. A trace is indicated by an X and accumulation by a Y punch.
49-52	Daily Total	Flagged for error when these data are omitted from the HPD #2 or keyed in the #1. When entered, the field is fully keyed and errors are indicated for blank columns. Trace is BBBX. Data are flagged for 000X, BOOX, etc. The daily total is checked with the values in cols. 13-48 of the HPD #1 & #2 cards. If the values do not agree it is indicated as a "cross-foot error" and the amount of error is shown. The cross-foot does not function if there are illegal punches.
53-56	Monthly Total	Keyed in the #2 card of the last calendar day of the month. This datum is listed at the bottom of the edit as card total. It is compared to the computed total taken from all daily #2 cards. If the totals are the same, the word "agree" is printed and, if not, the word "error" appears. If the total is omitted from the last #2 card, the card total is blank and error indicated.

C. Checking Procedure - Extreme Precipitation

1. The remaining data are contained in the HPD #4 card image. This card contains no date in cols. 10-11, and cols. 13-56 are blank. The card is listed on the edit below the last day of the month.

Column	Data	Edit Check
57-60	Greatest pcpn. in 24 hours	The value is checked for illegal punching.

Column	Data	Edit Check
61-65	Date of 24 hour amount	Col. 61 is keyed zero or X. Other values are flagged. When the value in 57-60 is 0000 these cols. will be blank and are flagged if not. The field is fully keyed if there is a value for 57-58 and listed as an "error pcpn. date" if miskeyed.
66-68	Greatest 24 hr. snowfall	Datum is keyed the same as the hourly pcpn. and has the same error check.
69-73	Date of 24 hr. snowfall	Same check as in cols. 61-65, with 69 keyed 0 or X with data in cols. 66-68.
74-75	Greatest snow depth	Zero is keyed for no snow. 2" = 02; 110 = X/10. Note: The snow depth is keyed in two cols., but prints to three places. This is to accommodate the overpunching for values greater than "99."
76-78	Date of snow depth	These cols. are blank if the value in cols. 74-75 is 00. 76 is keyed X for + dates or zero. Other values are flagged.
79-80	None	HPD cards 1, 2, and 4 are blank in these fields.

2. The edit contains a "Computed High - 24 Hour Precipitation" with dates. This is a guide to checking this value on the LCD. There is no check by the computer between this value and the one keyed in the HPD #4 card.

D. Correction of HPD Data

Data contained in the HPD 1, 2, and 4 card forms are corrected by submitting to the computer a new card punched in its entirety containing the information to be updated.

E. Maximum Short Period Precipitation.

For each month, maximum precipitation is keyed as two records: 1 in col. 10 with data for 5, 10, 15, 20, 30 and 45-minute periods, and 2 in col. 10 with data for 60, 80, 100, 120, 150 and 180-minute periods. See page 31 for the keying format. Day and time entries designate the end of the time period in which the amount of precipitation occurred. Day and time are omitted when the amount is zero, trace, or missing.

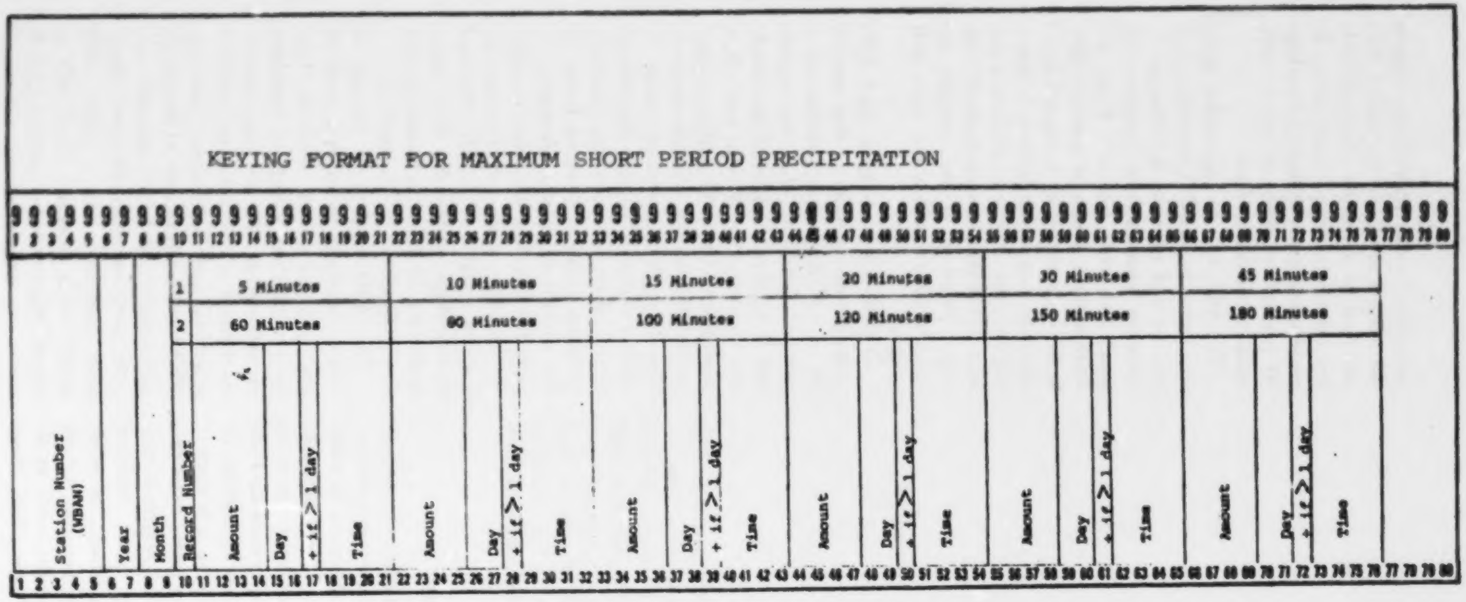
A computer edit program checks completeness and consistency of the data and produces an edit listing with flags indicating the deficiencies. The flags and associated deficiencies are as follows:

A = Record #1 missing
B = Record #2 missing
C = Month < 01 or > 12
D = Day < 01 or > 31
E = Hour < 00 or > 23
F = Minutes < 00 or > 59
R = Amount zero or trace
S = Missing "M"
T = Pcpn. ≥ 0.01 with day or time missing
W = 10.00 or greater

AA = 10 MIN > 2 X 5 MIN
AB = 15 MIN > 5 MIN + 10 MIN
AC = 20 MIN > 5 MIN + 15 MIN
AD = 20 MIN > 2 X 10 MIN
AE = 30 MIN > 10 MIN + 20 MIN
AF = 30 MIN > 2 X 15 MIN
AG = 45 MIN > 15 MIN + 30 MIN
AH = 80 MIN > 20 MIN + 60 MIN
AI = 100 MIN > 20 MIN + 80 MIN
AJ = 120 MIN > 20 MIN + 100 MIN
AK = 60 MIN > 2 X 30 MIN
AL = 150 MIN > 30 MIN + 120 MIN
AM = 120 MIN > 2 X 60 MIN
AN = 180 MIN > 60 MIN + 120 MIN

Corrections are keyed in the format shown on page 31, entering data for the time period involved only, for updating the tape. The updated tape produces printers copy for use in the CDNS Annual.

KEYING FORMAT FOR MAXIMUM SHORT PERIOD PRECIPITATION



Station Number (WBAN)	Year	Month	Record Number	Amount	Day	+ if > 1 day	Time	Amount	Day	+ if > 1 day	Time	Amount	Day	+ if > 1 day	Time	Amount	Day	+ if > 1 day	Time	Amount	Day	+ if > 1 day	Time	Amount	Day	+ if > 1 day	Time
			1	5 Minutes				10 Minutes				15 Minutes				20 Minutes				30 Minutes				45 Minutes			
			2	60 Minutes				90 Minutes				100 Minutes				120 Minutes				150 Minutes				180 Minutes			

AMOUNT B = BLANK

0 = XBBB
T = 0XBB } DAY WILL BE BLANK - TIME WILL BE BLANK
MISSING = BBBB - DAY MAY OR MAY NOT BE PRESENT

.10 = 0010
1.10 = 0110
10.10 = 1010 } DAY WILL ALWAYS BE PRESENT - TIME WILL ALWAYS BE PRESENT

MORE THAN 1 DAY, + = X

Validation, Compaction, and Analysis of Large Environmental Data Sets

By John Jalickee
Jerry Sullivan
Richard Rozett

EDS scientists have developed a technique which, among other benefits, allows them to compact a data set of 184,000 values into an equivalent data set of fewer than 6,000 values, while retaining 90 to 98 percent of the variability of the original data fields. Moreover, much of the remaining variability appears to be sensor noise.

Introduction

Large-scale environmental field experiments such as the Barbados Oceanographic and Meteorological Experiment (BOMEX), the International Field Year for the Great Lakes (IFYGL), and the GARP (Global Atmospheric Research Program) Atlantic Tropical Experiment (GATE) produce huge data sets and attendant large-scale problems in data validation, analysis, and synthesis. New and more sophisticated techniques are needed to extend and complement traditional methods when working with such large data sets.

The failure of conventional smoothing techniques to adequately remove noise from an IFYGL rawinsonde (atmospheric sounding) wind data set and still retain meaningful, though highly variable, natural fluctuations led the authors and other scientists of EDS' Center for Experiment Design and Data Analysis (CEDDA) to try a new method, called the asymptotic singular decomposition method, or ASD for short. The resulting computer program eliminated the noise and retained the essential data. It also greatly reduced the size of the original data base and, through intermediate graphics, provided a quick and efficient method of error detection, while isolating physical relationships and characteristic patterns.

The ASD Method

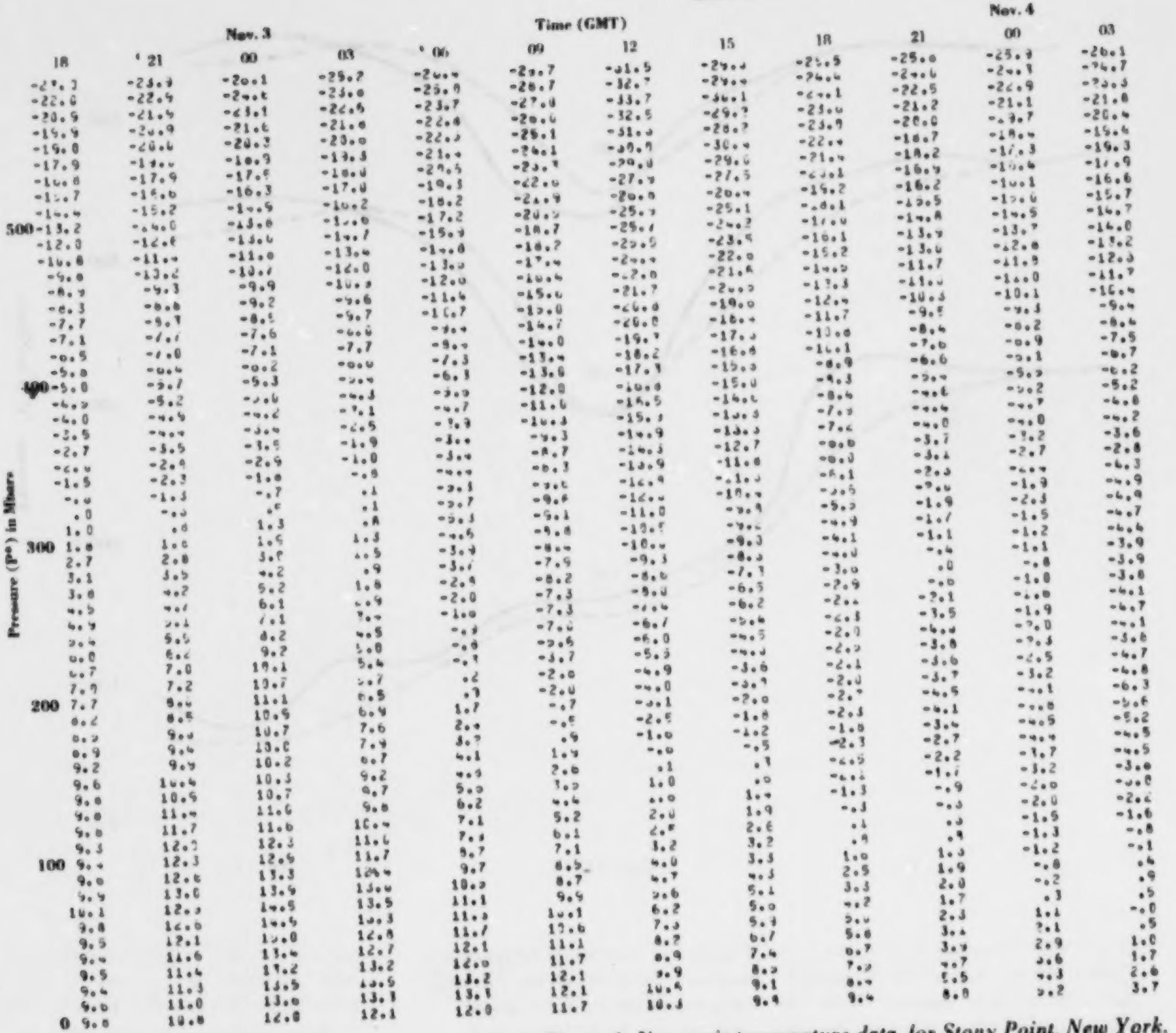
The idea behind this data decomposition technique is to extract meaningful information in the form of characteristic patterns. As an example, consider a meteorologist studying daily maximum temperature data for the east coast. Station by station, he observes that, in general, it is warmer in summer than in winter: from this he abstracts a typical seasonal variation. On the other hand, studying station-to-station variations, he notes that temperatures are generally colder in the north than in the south at almost any time of year. With these two characteristic variations (space and time) he can qualitatively explain the main features of the entire data set. And by retention of a relatively few significant temperature values he could quantitatively describe perhaps 90 percent of the east coast maximum temperature field.

The ASD data decomposition method adapted by CEDDA formalizes this process and provides a technique to calculate characteristic patterns for small and large data sets. Using the ASD method, dominant patterns within the data are easily extracted in an objective, repeatable fashion. In many respects, the science of ASD is much akin to the art of the caricaturist: the major features of the subject are quickly shown with a few sure, deft strokes.

CEDDA scientists have used ASD to reduce the quantity of data needed for a sufficient representation of a physical situation: often the equivalent data set is an order of magnitude smaller than the original one. Data generated by the method also are used in calculations that require relatively noise-free data: random noise is smoothed out, while real discontinuities or sharp changes are relatively unchanged. An unexpected bonus of the method is its error-detection capabilities: keeping with the caricaturist analogy, distorted (erroneous) features stand out sharply. Physical relationships within the data, often buried by the volume of numbers, are also highlighted by the method.

The ASD method is related to other statistical techniques such as principal component analysis, Lorenz's[1] empirical orthogonal functions in meteorology, and the factor analysis method of psychologists, political scientists, and sociologists; however, ASD has the advantages of simplicity and accuracy. A factor analysis computer program might fill over a thousand punched cards, while ASD would use a hundred. And ASD is almost immune to computer roundoff error, an important consideration when large data sets are involved.

Reproduced from best available copy.



Time (GMT)

Pressure (P*) in Mbars	18	21	Nov. 3 00	03	06	09	12	15	18	21	Nov. 4 00	03
	-27.3	-23.4	-20.1	-25.7	-24.4	-23.7	-31.5	-29.4	-25.5	-25.0	-25.3	-26.1
	-22.6	-22.5	-24.6	-23.8	-25.9	-26.7	-32.7	-29.4	-24.4	-24.6	-24.3	-24.7
	-20.9	-21.5	-23.1	-22.5	-23.7	-27.4	-33.7	-30.1	-24.1	-22.5	-22.9	-23.3
	-19.9	-20.9	-21.6	-21.8	-22.4	-26.4	-32.5	-29.7	-23.4	-21.2	-21.1	-21.8
	-19.0	-20.6	-20.3	-20.0	-22.3	-25.1	-31.4	-28.7	-23.9	-20.0	-19.7	-20.4
	-17.9	-19.4	-18.9	-19.3	-21.4	-24.1	-30.9	-30.4	-22.4	-18.7	-18.4	-19.4
	-16.8	-17.9	-17.9	-18.4	-20.5	-23.9	-29.0	-29.6	-21.4	-18.2	-17.3	-19.3
	-15.7	-16.6	-16.3	-17.8	-19.3	-22.6	-27.9	-27.5	-20.1	-16.9	-16.4	-17.9
	-14.4	-15.2	-14.5	-16.2	-18.2	-21.9	-26.8	-20.4	-19.2	-16.2	-16.1	-16.6
500	-13.2	-14.0	-13.8	-14.0	-17.2	-20.5	-25.9	-25.1	-18.1	-15.5	-15.6	-15.7
	-12.0	-12.8	-13.6	-14.7	-15.9	-18.7	-25.7	-24.2	-17.4	-14.8	-14.5	-14.7
	-10.8	-11.4	-11.0	-13.4	-14.8	-18.7	-25.5	-23.5	-16.1	-13.9	-13.7	-14.0
	-9.0	-10.2	-10.7	-12.0	-13.3	-17.4	-24.4	-22.0	-15.2	-13.6	-12.8	-13.2
	-8.9	-9.3	-9.9	-10.3	-12.0	-16.4	-22.8	-21.8	-14.3	-11.7	-11.9	-12.3
	-8.3	-8.8	-9.2	-9.6	-11.6	-15.6	-21.7	-20.3	-13.3	-11.3	-11.0	-11.7
	-7.7	-8.7	-8.5	-9.7	-10.7	-15.0	-20.0	-19.0	-12.4	-10.3	-10.1	-10.4
	-7.1	-7.7	-7.6	-8.6	-9.4	-14.7	-20.8	-18.4	-11.7	-9.5	-9.3	-9.4
	-6.5	-7.0	-7.1	-7.7	-8.4	-14.0	-19.7	-17.3	-10.8	-8.4	-8.2	-8.4
	-5.8	-6.4	-6.2	-6.4	-7.3	-13.4	-18.2	-16.8	-10.1	-7.6	-6.9	-7.5
400	-5.0	-5.7	-5.3	-5.4	-6.3	-13.6	-17.9	-15.3	-8.9	-6.6	-6.1	-6.7
	-4.9	-5.2	-5.6	-4.3	-5.9	-12.8	-16.8	-15.0	-9.3	-5.4	-5.3	-6.2
	-4.0	-4.5	-4.2	-3.1	-4.7	-11.6	-16.5	-14.6	-8.4	-4.6	-5.2	-5.2
	-3.5	-4.4	-3.4	-2.5	-3.9	-10.3	-15.1	-14.3	-7.3	-4.4	-4.7	-4.8
	-2.7	-3.5	-1.9	-1.9	-3.4	-9.3	-14.9	-13.3	-7.2	-4.0	-4.0	-4.2
	-2.4	-2.9	-2.9	-1.0	-3.4	-8.7	-14.3	-12.7	-6.0	-3.7	-3.2	-3.6
	-1.5	-2.3	-1.8	-.8	-4.4	-8.3	-13.9	-11.8	-6.3	-3.1	-2.7	-2.8
	-.4	-1.3	-.7	.1	-3.1	-9.6	-12.9	-11.3	-6.1	-2.3	-2.4	-4.3
	.0	-.3	.6	.1	-3.7	-9.6	-12.4	-10.4	-5.5	-2.4	-1.9	-4.9
	1.0	.6	1.3	.8	-3.3	-9.1	-11.0	-9.9	-5.3	-1.9	-2.3	-4.9
300	1.8	1.6	1.5	1.3	-4.5	-9.8	-10.5	-9.6	-4.9	-1.7	-1.5	-4.7
	2.7	2.8	3.8	1.5	-3.9	-8.4	-10.4	-9.3	-4.1	-1.1	-1.2	-4.4
	3.1	3.5	4.2	.9	-3.7	-7.9	-9.3	-8.3	-4.3	-.4	-1.1	-3.9
	3.8	4.2	5.2	1.8	-2.9	-8.2	-8.6	-7.3	-3.0	.0	-.8	-3.9
	4.5	4.7	6.1	2.9	-2.0	-7.3	-9.0	-6.5	-2.9	-.0	-1.0	-3.8
	4.9	5.1	7.1	3.4	-1.0	-7.3	-7.4	-6.2	-2.4	-2.1	-1.0	-4.1
	5.4	5.5	8.2	4.5	-.9	-7.6	-6.7	-5.6	-2.1	-3.5	-1.9	-4.7
	6.0	6.2	9.2	5.0	-.8	-5.6	-6.0	-4.5	-2.0	-4.3	-2.0	-4.1
	6.7	7.0	10.1	5.4	-.1	-3.7	-5.3	-4.4	-2.3	-3.8	-2.3	-3.8
	7.9	7.2	10.7	5.7	.2	-2.0	-4.9	-3.6	-2.1	-3.6	-2.5	-4.7
200	7.7	8.4	11.1	6.5	.3	-2.0	-4.0	-3.9	-2.0	-3.7	-3.2	-4.8
	8.2	8.5	10.9	6.9	1.7	-.7	-3.1	-2.0	-2.4	-4.5	-4.1	-6.3
	8.5	9.3	10.7	7.6	2.4	-.5	-2.5	-1.8	-2.3	-4.1	-4.0	-5.6
	8.9	9.4	10.6	7.4	3.7	.9	-1.0	-1.2	-1.8	-3.4	-4.5	-5.2
	9.2	9.9	10.2	8.7	4.1	1.9	-.0	-.5	-2.3	-2.7	-4.4	-4.5
	9.6	10.4	10.3	9.2	4.3	2.6	.1	.1	-2.5	-2.2	-3.7	-4.5
	9.8	10.9	10.7	9.7	5.0	3.5	1.0	.0	-2.4	-1.7	-3.2	-3.0
	9.8	11.4	11.6	9.8	6.2	4.4	1.0	1.4	-1.3	-.9	-2.0	-3.0
	9.8	11.7	11.6	10.4	7.1	5.2	2.0	1.9	-.3	-.3	-2.0	-2.2
	9.3	12.3	12.3	11.6	7.4	6.1	2.5	2.6	.1	.3	-1.5	-1.6
100	9.4	12.3	12.9	11.7	9.7	7.1	3.2	3.2	.9	.9	-1.3	-.8
	9.0	12.6	13.3	12.4	9.7	8.5	4.0	3.3	1.0	1.3	-1.2	-.1
	9.9	13.0	13.9	13.4	10.3	8.7	4.7	4.3	2.5	1.9	-.8	.4
	10.1	12.3	14.5	13.5	11.1	9.9	5.6	5.1	3.3	2.0	-.2	.9
	9.8	12.6	14.5	14.3	11.3	10.1	6.2	5.0	4.2	1.7	.3	.5
	9.5	12.1	15.0	12.8	11.7	10.6	7.3	5.3	5.4	2.3	1.1	-.0
	9.4	11.6	13.4	12.7	12.1	11.1	8.2	6.7	5.8	3.4	2.1	.5
	9.5	11.4	13.2	13.2	12.0	11.7	8.9	7.4	6.7	3.9	2.9	1.0
	9.4	11.3	13.5	14.5	13.2	12.1	9.9	8.3	7.3	4.7	3.6	1.7
	9.0	11.0	13.0	13.3	13.1	12.1	10.8	9.1	8.4	5.5	4.3	2.6
0	9.0	10.8	12.8	12.1	12.0	11.7	10.3	9.9	9.4	8.0	5.2	3.7

Figure 1. Upper-air temperature data for Stony Point, New York.

Data Compaction

Data from IFYGL for 1972-73 provide some vivid illustrations of the benefits of ASD applications. To demonstrate the data-compacting capabilities of the ASD method (plus the method itself), consider 12 successive IFYGL rawinsonde launches from station Stony Pt., N.Y., for the period 1800 GMT Nov. 2, 1972, to 0300 Nov. 4, 1972 (fig. 1). Temperature values are given for each 10-mbar pressure level, so that up to the 590-mbar level we have 12 × 60 = 720 values. (The pressure variable used in all figures is P^*, the difference between surface pressure and observed pressure, i.e., $P^* = P_{surface} - P$.) The particular time period was chosen because a sharp upper-air trough was passing over Lake Ontario, producing the characteristic temperature variations represented by the solid lines in figure 2.

The object of ASD application in this instance is to replace the 12 columns of 60 numbers with 1 column of 60 numbers and 1 row of 12 numbers, as in figure 3. In the latter illustration, the column represents the pressure dependence of the temperature soundings, while the row represents the time variation. To obtain the 350-mbar temperature for 0000 GMT on November 3, one would multiply the 36th number from the bottom of the column by the 3d number of the row (as shown in fig. 3), or,

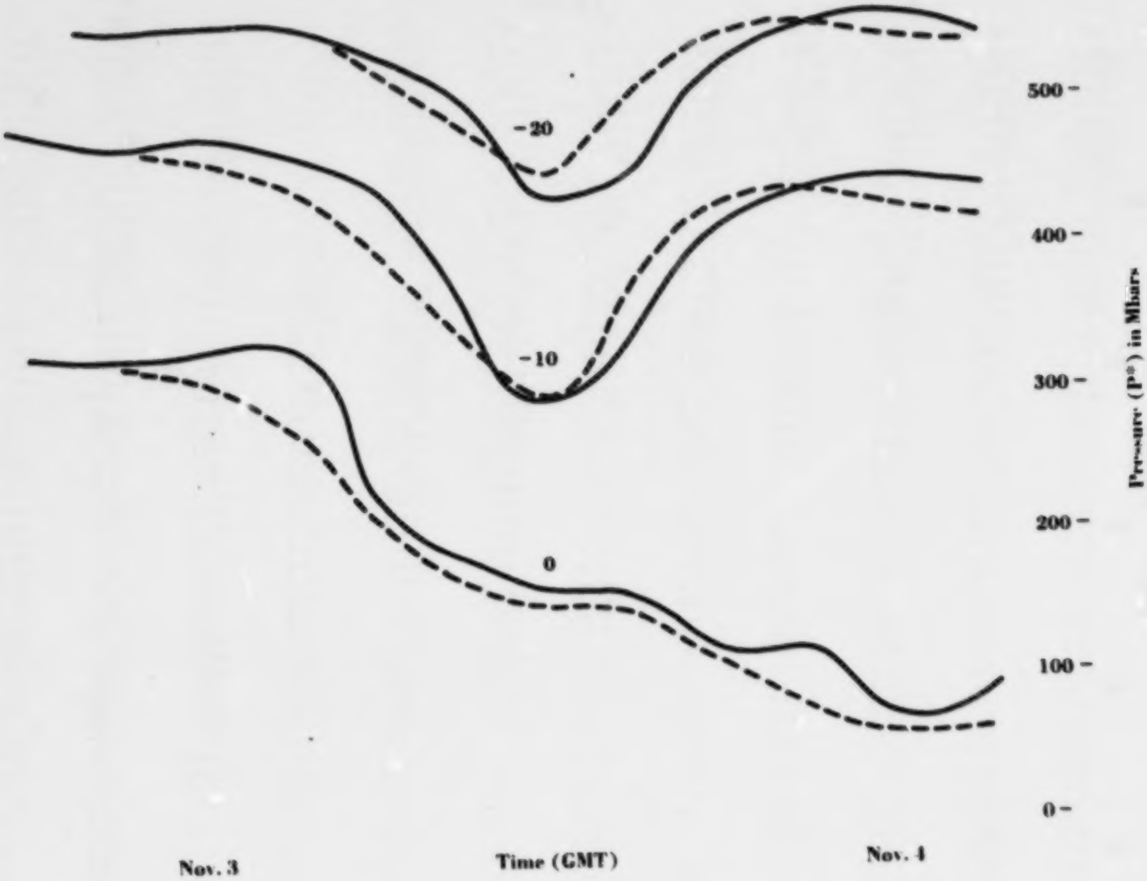


Figure 2. Time-height temperature analyses for Stony Point. The solid lines are based on the original data set, the dashed lines on a reconstituted data set.

to get the 150-mbar temperature at 1800 GMT on November 3, multiply the 16th column number from the bottom by the 9th number in the row.

Where did the column and row come from? Any column and row of numbers can be multiplied together to generate a temperature field. The best choice is one that minimizes the sum of squared differences between the generated field and the original field. In practice, the ASD computer program begins with a trial column and row, then generates successive values until there is no further minimization of differences between the two temperature fields.

In the example at hand, the original 720 numbers have been replaced by 60 + 12 = 72 numbers, a 10-fold reduction. The new field generated by the row and column explains approximately 90 percent of the variation about the mean of the original field. The ASD method now may be used again to describe the residuals of the original field minus the first generated field, producing another row and column. Usually, about 98 percent of the original temperature field variation is covered by three rows and columns. The broken lines in figure 2 show a reconstituted temperature field using three rows and columns.

Overall, CEDDA scientists were able to compact 60 levels of temperature, humidity, and wind values from 768 IFYGL upper-air soundings (6 stations, 128 launches each), a total of 184,000 values, into an equivalent data set containing fewer than 6,000 values. From 90 to 98 percent of the characteristics of the original fields are retained, and much of the unexpected variability appears to be sensor noise.

Error Checking

Figure 4 illustrates ASD's error-detection capability. Obviously, the sounding for station 2 differs greatly from the soundings for the other five stations. Figure 5 shows the time components corresponding to the pressure component of figure 4. Once again, a strong anomaly (circled values) shows up. The six soundings indicated were checked and did prove to be erroneous. Thus, a 10-second scan of these two ASD graphs isolated an error that previously had escaped detection.

Physical Relationships

Three station pairings stand out clearly in the lower levels of the soundings shown in figure 6. These station pairs—1-2, 3-6, and 4-5—are geographically related. Stations 4-5 are on the western end of Lake Ontario, 3-6 on the middle shoreline, and 1-2 on the eastern end. Figure 7, a plot of the corresponding time component, shows that the effect is most pronounced for launches number 20 through 27. A detailed check of the soundings from all stations for this period revealed a large east-west wind velocity gradient which varied from 2 m/s in the west to 6 m/s in the middle to 14 m/s in the east.

Other Uses

With ASD, new data can be compared quickly with older data obtained by the same measuring system. Drastic differences in the ASD plots will suggest instrument drift and/or mistaken assumptions about experimental background conditions. The same approach can be used where different types of instruments are supposedly measuring the same physical phenomenon. This type of application allows CEDDA scientists to study the very large data sets associated with ecosystems and often, through simultaneous analysis of many different kinds of variables, uncover hidden interactions.

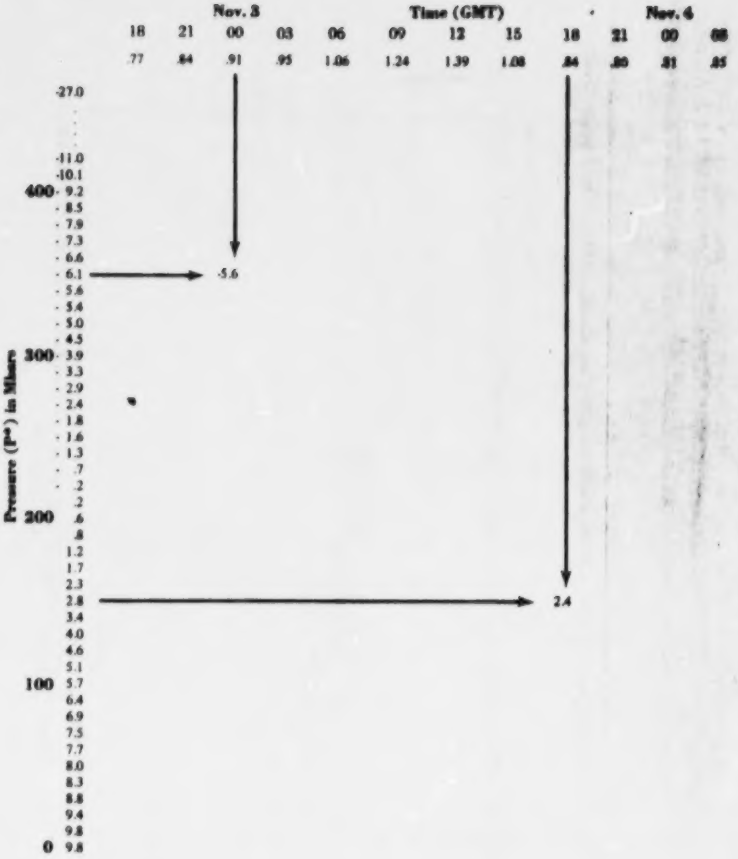


Figure 3. An illustration of the ASD data compaction technique. The single column of 60 numbers and single row of 12 numbers replace the 12 columns of data appearing in figure 1, yet retain approximately 90% of the details of the original data set.

Modeling and Experiment Design

CEDDA scientists are pursuing other potential applications of the ASD method, including its use in modeling and experiment design. The characteristic patterns obtained provide important clues as to the physical realities underlying the data. We hope that the pattern-detection capabilities of the ASD method may lead to an empirical, data-oriented form of system modeling.

Another promising path leads towards the economical design of field experiments and data collection systems, based on characteristic patterns derived from preliminary survey data. Much redundant data and information are often collected in large-scale field experiments. If the redundancies could be eliminated, all subsequent data collection, processing, analyses, archival, and dissemination activities would be greatly simplified and more cost-effective. The ASD method, by highlighting significant patterns of preliminary survey data sets, could suggest which data contribute most to the definition of the patterns, and which are dispensable.

Reference

[1] Lorenz, E. H., *Empirical Orthogonal Functions and Statistical Weather Prediction*, MIT Department of Meteorology, Science Report No. 1, 1956.

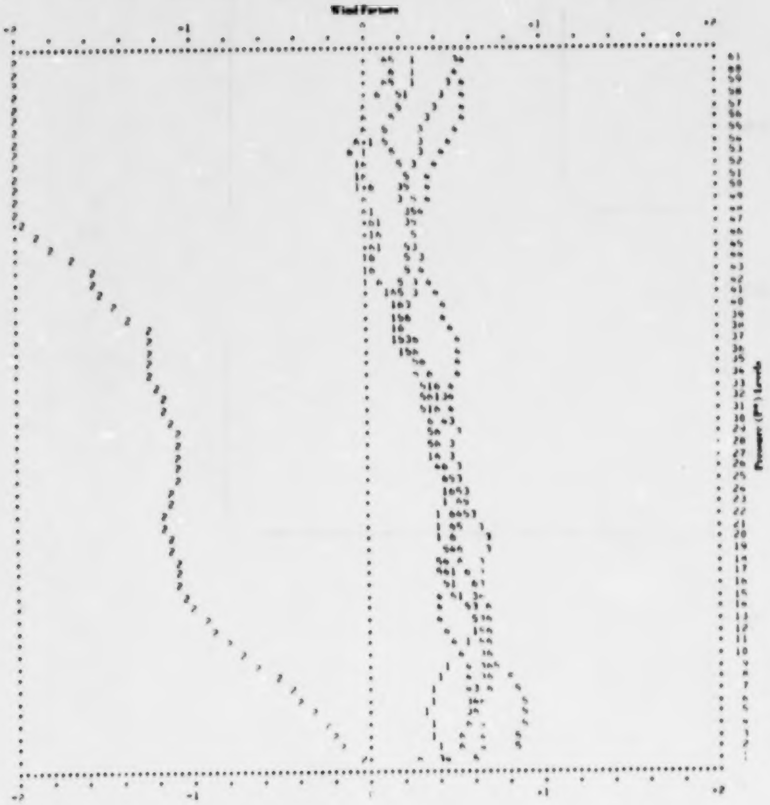


Figure 4. Composite printout of U-components of the wind for 48 upper-air soundings taken at each of six IFYGL observation stations.

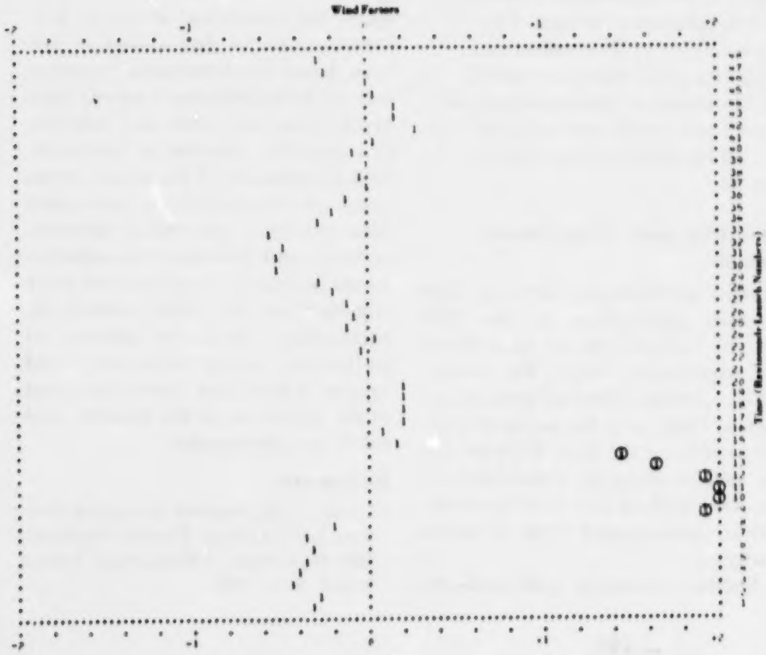


Figure 5. Time analysis of data from figure 4 isolates six anomalous soundings (circled).

Figure 6. Composite of V-components of the wind for 48 upper-air soundings taken at each of six IFYGL observation stations.

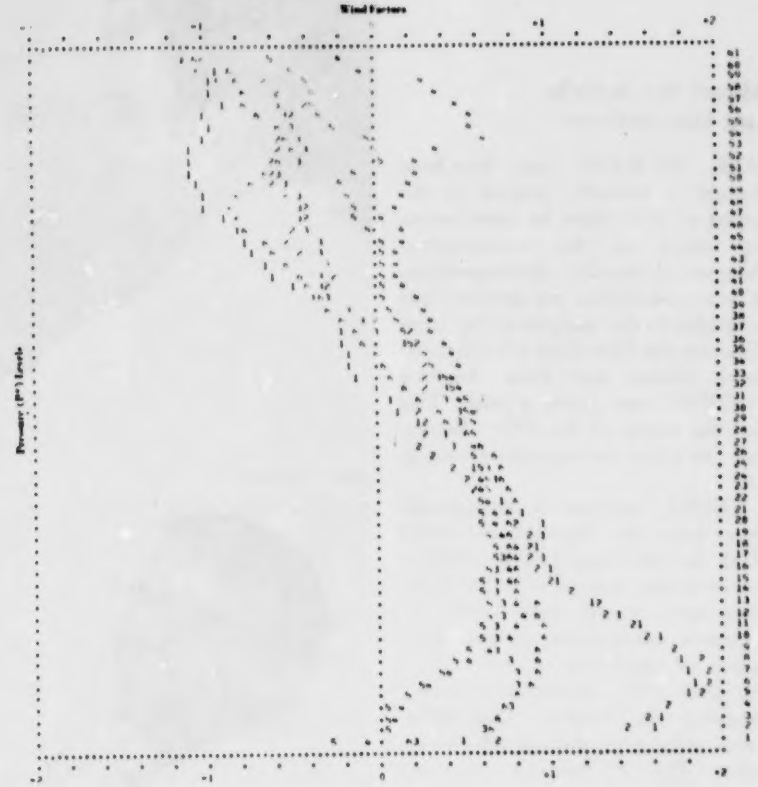
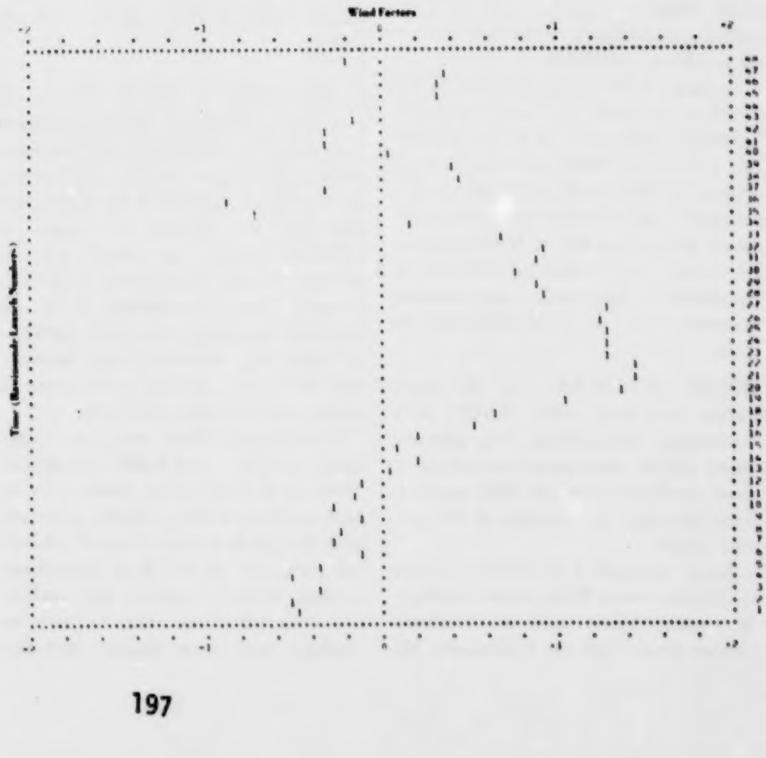


Figure 7. Time analysis of data from figure 6 indicates that the pairing pattern is most pronounced in soundings 20-27.



Jack Jalickee

Dick Rozett

Jerry Sullivan

About the Article and the Authors

JACK JALICKEE was thumbing through a scientific journal in the spring of 1973 when he came across an article on the mathematical theorem of singular decomposition. It was evident that the theorem was adaptable to the analysis of the large data sets the EDS Center for Experiment Design and Data Analysis (CEDDA) was working with. This was the origin of the ASD (Asymptotic Singular Decomposition) method.

CEDDA analysis of atmospheric data from the International Field Year for the Great Lakes (IFYGL) began in the autumn of 1974. Problems arose almost immediately. Divergence calculations derived from upper air winds did not make physical sense. (The calculation is a very sensitive one, involving small differences of large numbers which contain noise.) The data themselves appeared reasonable and consistent with observed weather conditions, which were highly variable. Traditional analysis techniques could not resolve the problem; ASD did.

A native of Washington, D.C., Jack Jalickee worked his way through Catholic University (in D.C.), receiving a B.A. in 1962, and a Doctor's degree in 1966, both in Physics. Subsequently, he worked as a research associate and teacher at Northwestern University in Evanston, Illinois. A Presidential Internship appointment brought him to EDS/CEDDA in 1972.

JERRY SULLIVAN was the man having problems with IFYGL data divergence calculations. His inhouse paper on the subsequent resolution of those problems through ASD applications provided the nucleus of the current article.

Jerry received a Bachelor's degree in Physics from Holy Cross College, Worcester, Mass., and his Doctor's Degree from Catholic University. He joined EDS/CEDDA in the fall of 1970.

Fr. RICHARD ROZETT, S.J., is on a year's sabbatical from Fordham University in New York. His previous work and interest in the application of statistical techniques to large data sets led Fr. Rozett to come to CEDDA, where he heads up its MESA (Marine Ecosystem Analysis) Project. Since September 1974, he has been working with Jack Jalickee in collecting, devising, and developing ASD and similar techniques to analyze ecosystems data sets.

Ecosystems data sets are very large, complex, and highly redundant. They include physical measurements such as temperature, depth, pressure, and the particle size of sand; chemical measures of oil, lead, phosphate, acidity, salinity, nitrate, and carbonate concentrations—not to mention garbage and sewer sludge; and biological measurements such as the number of barnacles per square meter, or the percent of flounder with fin rot. ASD and similar techniques make it possible to massage the original data into a simpler, concentrated, and more meaningful data set.

Dick Rozett earned a B.S. degree in chemistry from Spring Hill College in Spring Hill, Alabama, a M.S. degree from St. Louis University, then studied chemical physics at Johns Hopkins in Baltimore, Md., where he received his Ph.D. in 1967.

Ordained a priest in 1962, Fr. Rozett was an Assistant Professor of Chemistry at Fordham from 1967 to 1972, when he was made an Associate Professor. He is the author of more than 30 scientific papers on chemistry and the statistical analysis of large data sets, and has participated in international scientific conferences in Leningrad, Lisbon, Kifissia (Greece), and Kyoto.

DATA VALIDATION FOR UPPER AIR SOUNDING DATA AND EMISSION INVENTORY DATA

by

J.H. Novak

Environmental Sciences Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

DATA VALIDATION FOR UPPER AIR SOUNDING DATA AND EMISSION INVENTORY DATA

J.H. Novak

A systematic approach to data validation requires that several steps be taken during the design of a validation scheme. For any set of data it is essential to be familiar enough with the data collection and data handling procedures to be able to locate all possible sources of error and to define a criteria for distinguishing good and bad data at those critical points. The next task is to determine which techniques can be used most effectively in error checking, and what course of action should be taken if an error is detected. Finally, after the validation scheme has been implemented the quality of the validated data should be assessed in some manner.

Therefore the first step in the validation of RAPS upper air data was to determine all possible sources of error in the data handling system. The upper air data consists of two types of observations, Pibals and Radiosondes.

A pibal is a pilot ballon which is filled with helium to an exact pressure in order to insure that it will rise with a known ascension rate when released into the atmosphere. An observer uses a mechanical device known as a theodolite to track the balloon by recording azimuth and elevation angles at 30 second intervals. These angles are then used to calculate wind speed and direction at various heights above ground. There are two possible sources of error during this phase of data collection. First, the observer may read the angles incorrectly during the sounding and second, transcription errors may occur when coding the data onto forms for keypunching.

The radiosonde is similar to a pibal in that it is also a balloon; however, a package of instrumentation containing various meteorological sensors is attached to the balloon which is tracked electronically instead of manually. In addition to the azimuth and elevation angles, pressure, temperature, and relative humidity readings are recorded. A variety of thermodynamic parameters can be determined from these measurements. There are several potential sources of error associated with the soundings : electronic difficulties, sensor malfunction, calibration, misinterpretation of the strip charts, interpolation of the adiabatic charts and transcription errors.

Once all possible sources of error have been determined and a range of good and bad data defined, various techniques can be chosen to search the data for possible errors. The upper air sounding network's(UASN) preliminary quality control program contained the following tests on the raw data:

1. Routine data checks - data was checked for completeness and compared with known data(e.g. station date and time vs a performance matrix, station # vs station height, balloon weight vs release time)
2. Consistency checks with alternate data source(e.g. wind data vs station log books, doubtful data vs weather maps and recording barograph)

3. Intra-station checks with previous and following soundings.
4. Inter-station checks with simultaneous soundings.
5. Checks with known meteorological relationships(e.g. comparison of temperature and relative humidity with adiabatic charts, shape of the pressure-altitude curve).

The actual key punching of the data forms introduces another source of error. But at this point the data checks can be computerized, so that all data will routinely undergo the same tests. The UASN data validation programs test the data for order, range, missing values, station height, and special conditions such as calms or wind speeds greater than 40 meters/second. Again, additional checks can be performed on the radiosonde data when special relationships exist(e.g. inverse relationship between pressure and time). The advantage of computerized error checking is that the entire data set can be objectively evaluated.

Once the known data errors have been flagged and corrected, the next step is to archive the data. During this phase both printouts and printer plots are produced in order to provide additional information to be used for error detection. The printer plots of speed, direction, temperature, dew point and atmospheric pressure can quickly be scanned for remaining inconsistencies.

In an effort to further validate the UASN radiosonde data , dew point, relative humidity and vapor pressure calculated at sites 141 and 142 were compared with corresponding data recorded by the national weather service at Lambert Field. Correlation and incidence matrices were also calculated for the same three parameters on a seasonal basis. Both types of preliminary analysis proved very effective in isolating some remaining data problems.

The final quality assurance effort produced Calcomp plots of wind speed, direction, temperature, potential wet bulb temperature and mixing ratio for each of the 5,717 UASN radiosondes. Each of these plots were scanned for data errors and used to determine mixing depths for the St. Louis area.

In summary, the important concepts to be derived from the previous discussion of validation techniques used with the RAPS Upper Air Sounding Network data are:

1) Determination of all possible sources of error in the collection and data handling.
2) Use of alternate sources of data for consistency checks.
3) Use of intra and inter station comparisons.
4) Use of known relationships(meteorological in this case) for comparisons.
5) Completeness and objectivity of computerized comparisons.
6) Use of preliminary analysis routines in error detection.
7) Use of computer graphics.

The second topic for discussion is the validation of the RAPS emission inventory. The main objective of the RAPS program is to provide a body of data (emissions, meteorological, air quality, etc.) which could be used to develop, improve and validate air quality simulation models. The first priority is to determine what accuracy is required in any data base to be able to achieve the objectives of RAPS and secondly, what accuracy, precision, and bias currently exists in the RAPS emission inventory. The answers to these questions are too complex to be addressed in this paper, but they are essential to the design of a good validation scheme; therefore I have included(as references) a list of papers which discuss this important question of accuracy in detail. Thus, from this point I will limit the discussion to the procedures that were chosen to verify the accuracy of the acquired and estimated data.

The RAPS emission inventory is composed of three separate data bases: (1) point, (2) area, and (3) line source. The choice of validation technique depends on the amount and form of the data in each data base. The point source data base contains hourly, daily, monthly and annual raw process data ; no emissions are stored in the data base. The methodologies used to calculate emissions and determine temporal resolution are applied at data retrieval time. In contrast, the area and line source data bases contain annual emissions. The methodologies used to calculate emissions have already been applied before the data was entered into the data base. Temporal apportionment is accomplished through the retrieval software. As usual, checks must be performed on raw data at their entrance into the data handling system . For the area source data base, this implies checking the raw data inputs to the methodology programs. There are seven source categories for the area source inventory - river vessels, fugitive dust, highways, railroads, stationary residential and commercial sources, off-highway mobile sources, and stationary industrial sources and airports. The software for these source categories was developed by several different contractors and therefore must be reviewed independently. Area source data is mainly checked for internal consistency within each grid. Parameters such as population, number of homes, amount of water area per grid, agricultural acreage etc, are compared with each other in terms of overall land use per grid. Typical errors that were found include a 1 KM square grid which contained over 2 million acres of tilled farm land and a grid with population of 180 and only 11 single family homes. Calcomp graphics was heavily used in the validation of line source data. Line sources and associated characteristics such as average daily traffic, functional class, etc. were plotted on gridded maps to the same scale as county roads and DOT maps. Overlaying these maps provided an excellent means of checking the raw line source data.

In contrast with the area source data base, the point source data base contains all the raw data for emission calculations. Because the raw point data includes temporally-distributed process data for the entire study period in contrast to annual county statistics for area data, the amount and type of point data must be taken into account when choosing a technique for raw data validation. Parameters which apply at the stack level and therefore do not have a temporal association can be manually verified against original plant data. These parameters include stack and fuel characteristics, operating patterns, stack test data, and applicability of the SCC to a given stack. And because of the small amount, monthly process data was verified manually.

In order to perform a reasonable check on the remaining data a random selection of representative sources were chosen. The prime determinants in the selection of test sources were the method of emission calculation and the time interval of reporting the data. One source from each combination of these two factors was chosen to insure that all paths in the software would be exercised. The following tests were performed on the selected sources: 1) manual verification of process data, 2) verification of diurnal, weekly and/or seasonal variations, 3) hourly and annual retrievals. Computer software was developed to check all process data in the point data base for consistency and continuity.

Finally, all test software runs were compared with hand calculations and the retrieval programs themselves were compared with the documented methodologies.

In summary, the important concepts to be derived from the above discussion of the validation of RAPS emission inventory data are:

1) Preliminary determination of required accuracy.
2) Analysis of current accuracy.
3) Selection of validation techniques by:
    a) amount of data
    b) form of data
    c) availability of supporting data
    d) significance of data to the overall accuracy
    e) availability of time and personnel

REFERENCES

Kock, R.C. et al, "Validation and Sensitivity Analysis of the Gaussian Plume Multiple-Source Urban Diffusion Model", NTIS Publication Number PB-206951, Geomet Inc., Rockville, Maryland(1971).

Ditto, F.H. et al, "Weighted Sensitivity Analysis of Emission Data", Final Report, EPA Contract # 68-01-0398(1973)

Littman, F.E., S. Rubin, K.T. Semrau, and W.F, Dabberdt, "A Regional Air Pollution Study(RAPS) Preliminary Emission Inventory", SRI Project 2579 Final Report, EPA Contract #68-02-1026 (1974).

Gibbs, L.L., C.E. Zimmer, and J.M. Zoller, "Source Inventory and Emission Factor Analysis", Volumes I and II, Final Report, EPA Contract # 68-02-1350 (September 1974).

Ruff, R. E., P. B. Simmon, "Evaluation of Emission Inventory Methodologies for the RAPS Program", SRI Project 4331, Final Report, EPA Contract # 68-02-2047 ((1977).

VALIDATION OF BIOMEDICAL DATA THROUGH AN ON-LINE COMPUTER SYSTEM

by

Larry D. Claxton

Health Effects Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

VALIDATION OF BIOMEDICAL DATA THROUGH AN ON-LINE COMPUTER SYSTEM

L.D. Claxton

INTRODUCTION

Within the biomedical disciplines there are a variety of testing procedures used routinely within many separate laboratories. Since health, research and regulatory decisions are being based upon the results from many laboratories, there is a basic need for assuring the quality of the data. In the area of microbial mutagenesis, the use of *Salmonella typhimurium* as an indicator organism for mutational events is employed by many laboratories across the country. The various procedures available are rapid, relatively simple, sensitive and are used in a variety of laboratory situations including private industry, government and university laboratories. Presently, a great deal of emphasis is placed upon these types of tests as prescreens for substances that may be human mutagens and potential carcinogens. Therefore, the use of a system involving *Salmonella typhimurium* could provide an excellent pilot study for methods involved in data validation. Data validation is used in this context to mean the process by which generated data is filtered and accepted or rejected by objective criteria. Likewise, computerization provides a potential means for systematically applying a predetermined set of objective criteria in a rapid non-biased manner. With the use of TSO (Time Sharing Option), portions of the data validation can be conducted during the performance of a biological test. This article will describe the design of a pilot system for the on-line computer assistance of testing

---

Also published separately as EPA-600/1-78-038, "Biomedical Data Validation Through an On-Line Computer System," May 1978.

protocols and data validation. The scientific protocols and initial computerization have been completed and the system will be tested in a laboratory situation in the near future by the National Institute of Environmental Health Sciences.

DESCRIPTION OF TEST:

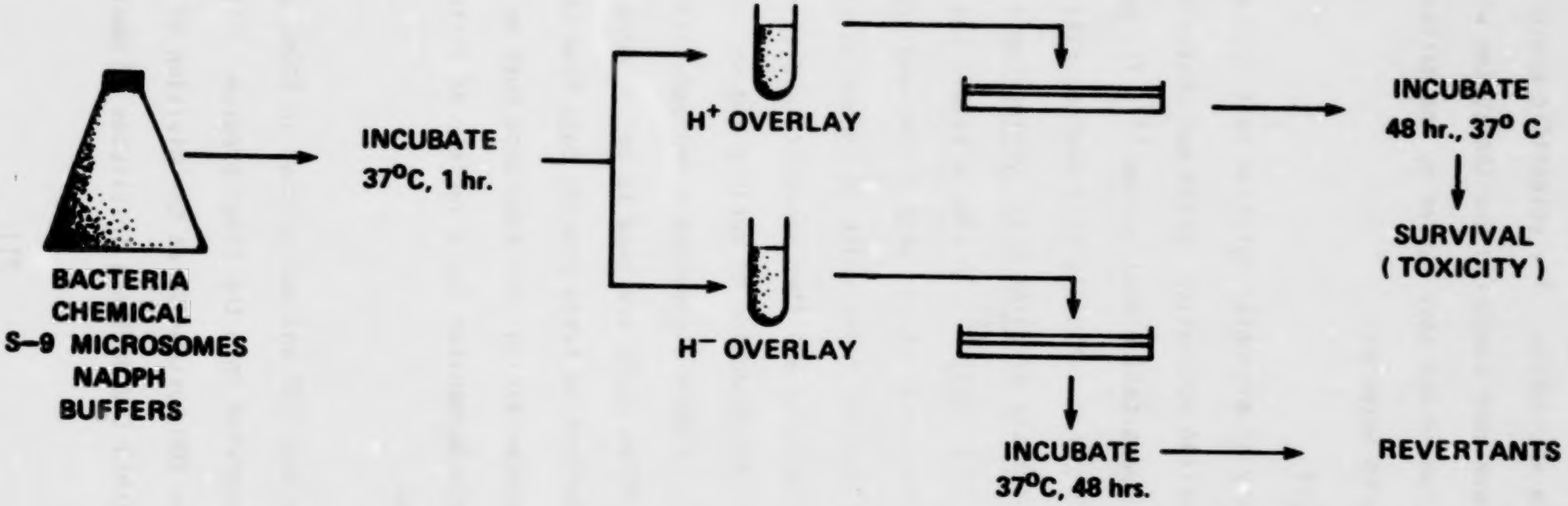
From a variety of microbial mutation test systems, the suspension test using a mammalian activation system was chosen because it is well defined and is a quantitative test system.(1) The more commonly used Ames plate incorporation method is only semiquantitative. We also chose to compare three strains of Salmonella typhimurium and a forward mutation strain of K-12 E. coli.(2) In simple terms, the test involves the combining of the bacterial strain with a compound and a mammalian activation system into an Erlenmeyer flask which is incubated at 37°C for 30 minutes to 2 hours. The bacteria are then separated and aliquots are plated on minimal media for the detection of mutants and on supplemented media for relative survival. Figure 1 provides a representation of the pilot test presently used. Pilot tests are used to define more appropriate testing conditions, and definitive tests provide data from which mutagenicity is judged. For complete testing, the substance must be tested in several strains of bacteria to monitor for a variety of different types of genetic alteration.

SYSTEMS OVERVIEW

This program uses TSO and was written in COBOL with some additional FORTRAN being integrated into the final program. All programming was accomplished on an IBM System/370 at the Division of Computer Research and Technology within the National Institutes of Health, Bethesda, Maryland.

Figure 1.

REPRESENTATION OF PILOT SUSPENSION TEST FOR MICROBIAL MUTATION


BACTERIA
CHEMICAL
S–9 MICROSOMES
NADPH
BUFFERS
INCUBATE
37°C, 1 hr.
H⁺ OVERLAY
H⁻ OVERLAY
INCUBATE
48 hr., 37° C
SURVIVAL
(TOXICITY)
INCUBATE
37°C, 48 hrs.
REVERTANTS

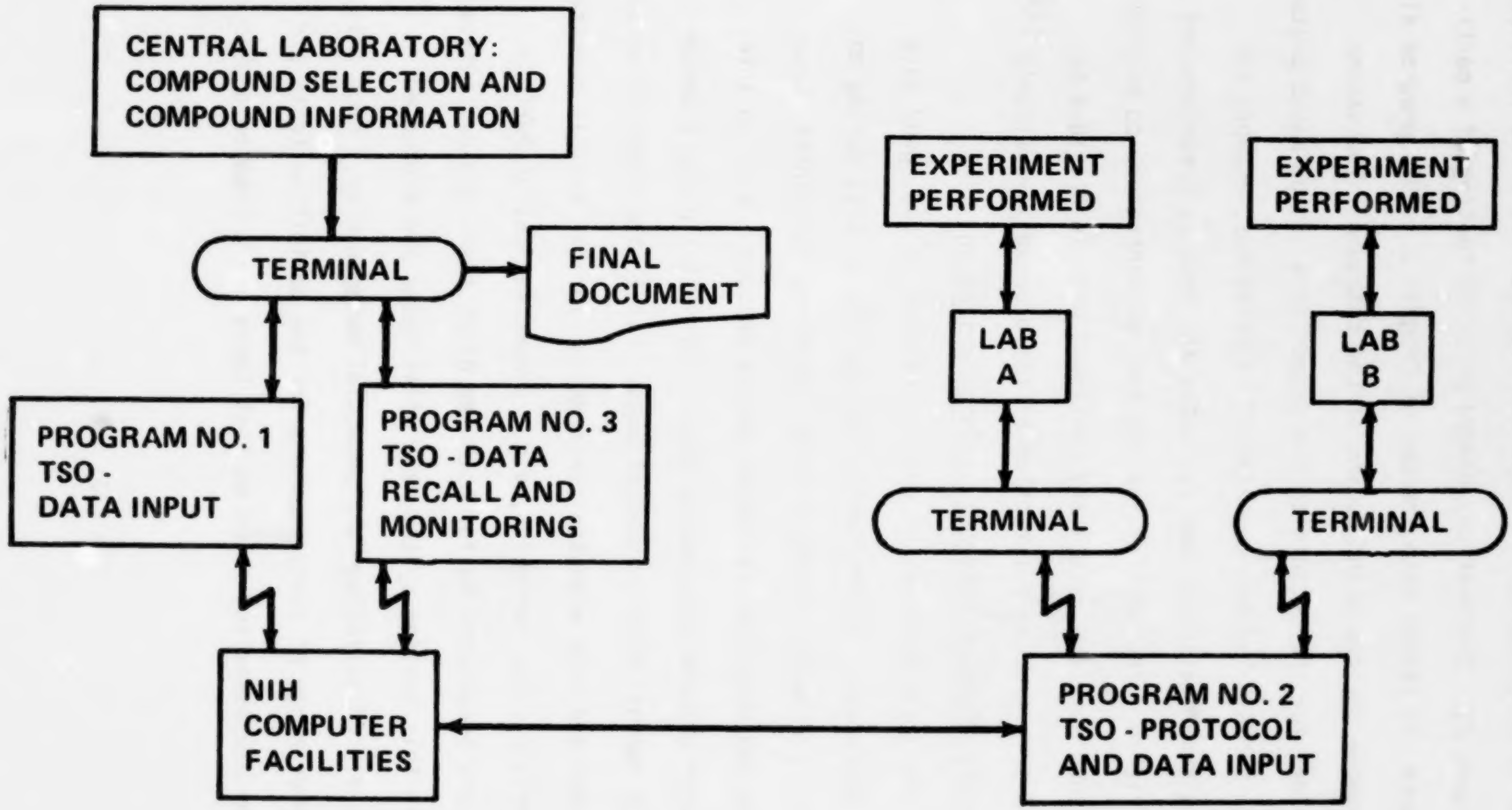
For ease of programming, the task was divided into three individual programs (Figure 2). Information, needed prior to testing of a particular substance, is stored with the use of Program 1. This program also supplies a number for the blind coding of the compound. The second program provides for the technician the proper form of the basic protocol, performs certain "within-experiment" calculations, accepts the input of data from the tests, and evaluates the test by predetermined objective criteria. The ability for the central laboratory to monitor the accomplished work and recall any pertinent data is provided by Program 3. A more precise description of the program is available.(3)

Quality Control Through Interactive Computerization

One of the basic premises of quality control is that good data yields good decisions. By monitoring the quality of data during an experiment and providing feedback to the technical personnel, both personal bias and technical variation can be reduced. With an interactive computer network this can be done. This pilot project demonstrates these capabilities in several ways. First, the compound to be tested is coded and only essential information for the test is provided. Secondly, certain other variables, e.g., concentrations of various components, are predetermined for both the pilot tests and definitive test. Within this testing system, two pilot tests are conducted to determine levels of toxicology and potential mutagenicity. From this data a narrower range of concentrations for the definitive tests are calculated by predetermined rules so that there are a limited number of

Figure 2.

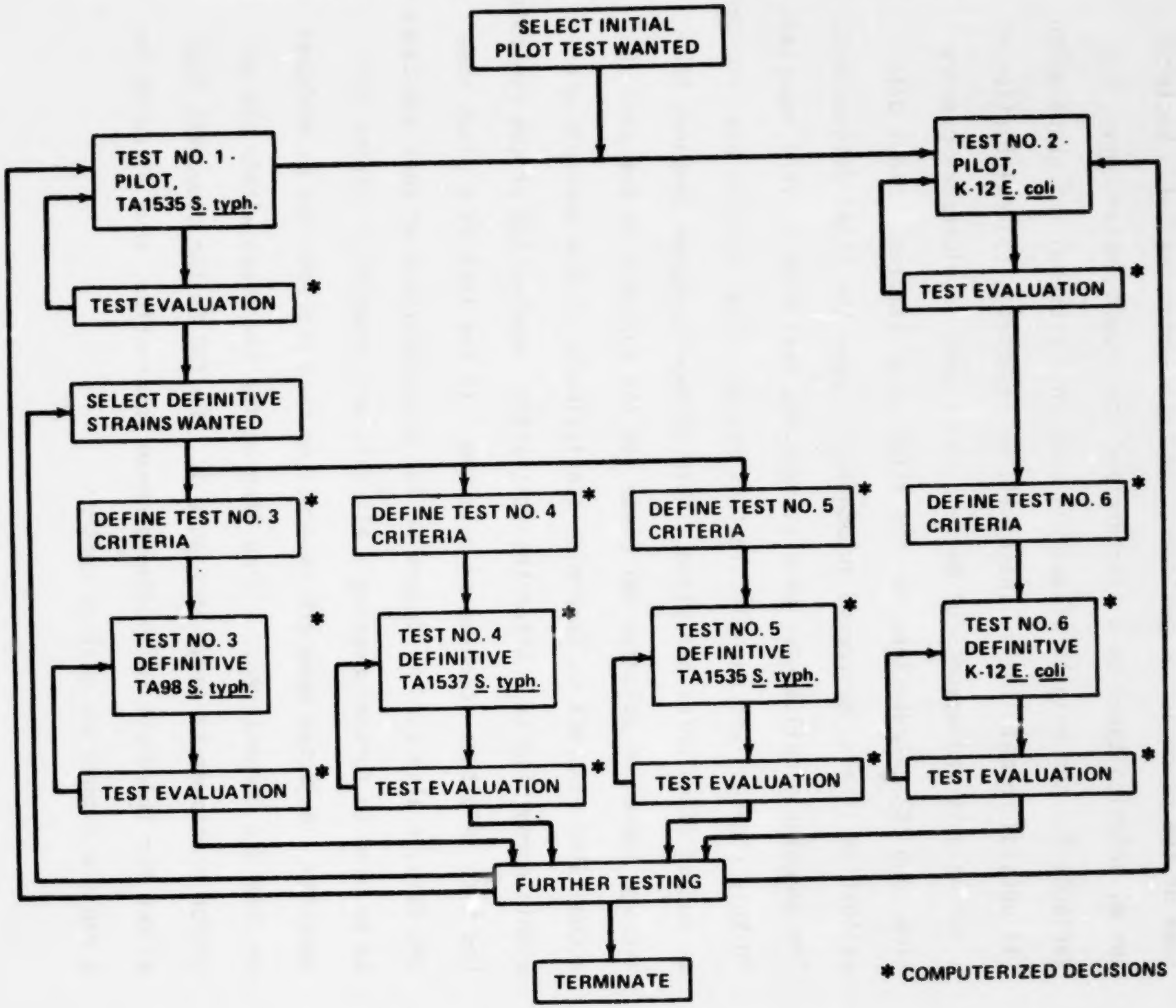
SYSTEM FOR COMPUTERIZED ASSISTANCE IN MICROBIAL MUTATION TESTING


CENTRAL LABORATORY: COMPOUND SELECTION AND COMPOUND INFORMATION
TERMINAL
FINAL DOCUMENT
PROGRAM NO. 1 TSO - DATA INPUT
PROGRAM NO. 3 TSO - DATA RECALL AND MONITORING
NIH COMPUTER FACILITIES
EXPERIMENT PERFORMED
LAB A
TERMINAL
EXPERIMENT PERFORMED
LAB B
TERMINAL
PROGRAM NO. 2 TSO - PROTOCOL AND DATA INPUT

definitive concentrations used across all laboratories. Next, the computer performs any needed calculations during the performing of a test thus lessening the occurrence of potential computational errors. Some of the calculations performed for this system are: (1) bacteria per ml solution based on a standardized spectrophotometer curve, (2) variance for the weights of animals used in microsomal S-9 preparation (if outside normal limits, these will be rejected), (3) calculation of liver weights and amounts of buffers to be used in microsome preparation, and (4) calculations for the dilution of samples. Final data validation is also performed automatically upon the final data output. The computer's ability for data storage and retrieval is very important in this regard. For example, in this system, final results are recorded as number of colonies per plate. This software program compares the average number of colonies per plate for the controls to the past 100 accumulated controls to determine statistically if the controls are within normal limits. After the statistical examination of the controls the test is either accepted or rejected. If the test is a pilot then the data is also used to determine the concentrations of test substance to be used in further testing. All data are, however, recorded permanently. Rejected data are recorded so that problems can be analyzed as they are encountered. A flow diagram for the areas within the decision processes is shown in Figure 3. The TSO is the component that allows for immediate technician/program interaction, thus allowing for a rapid and constant quality control.

Figure 3.

SUMMARY OF THE DECISION PROCESS FOR A SINGLE COMPOUND


SELECT INITIAL PILOT TEST WANTED
TEST NO. 1 - PILOT, TA1535 S. typh.
TEST NO. 2 - PILOT, K-12 E. coli
TEST EVALUATION *
TEST EVALUATION *
SELECT DEFINITIVE STRAINS WANTED
DEFINE TEST NO. 3 CRITERIA *
DEFINE TEST NO. 4 CRITERIA *
DEFINE TEST NO. 5 CRITERIA *
DEFINE TEST NO. 6 CRITERIA *
TEST NO. 3 DEFINITIVE TA98 S. typh. *
TEST NO. 4 DEFINITIVE TA1537 S. typh. *
TEST NO. 5 DEFINITIVE TA1535 S. typh. *
TEST NO. 6 DEFINITIVE K-12 E. coli *
TEST EVALUATION *
TEST EVALUATION *
TEST EVALUATION *
TEST EVALUATION *
FURTHER TESTING
TERMINATE
* COMPUTERIZED DECISIONS

This prototype system demonstrates that interactive computer programs can be used to effectively increase the quality control of rapid *in vitro* tests. However, it is also apparent that the more simple *in vitro* microbial mutagenesis tests such as spot tests and simple plate incorporation tests do not require such extensive computerization if well documented and detailed protocols are available. Since most *in vivo* mammalian systems have extended experimental time periods, the time sharing option would be of little benefit due to cost factors and experimental design. However, even with the more simple *in vitro* tests and mammalian cell culture tests, this system can serve as a model for data storage and test evaluation for the purpose of quality control.

This paper was extracted from an EPA report(4) which is available through the National Technical Information Service, Springfield, Virginia 22161.

REFERENCES

1. Frantz, C. N. and Malling, H. U. 1975. The Quantitative Microsomal Mutagenesis Assay Method. Mutation Research 31:365-380.
2. Mohn, Georges, Ellenberger, J. and McGregor, D. 1974. Development of Mutagenicity Tests Using *Escherichia coli* K-12 As Indicator Organism. Mutation Research 25:187-196.
3. Claxton, Larry and Baxter, Richard. 1978. The Computer Assisted Bacterial Test for Mutagenesis. Mutation Research (In Press).
4. Claxton, L. Biomedical Data Validation Through An On-Line Computer System. EPA-600/1-78-038, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, May 1978. 10 pp.

REGIONAL VALIDATION OF STATE AND LOCAL AIR POLLUTION DATA

by

Thomas H. Rose

Region IV
U.S. Environmental Protection Agency
Athens, Georgia 30605

REGIONAL VALIDATION OF STATE AND LOCAL AIR POLLUTION DATA

Thomas H. Rose

SUMMARY

Two types of data auditing are performed on state and local data in the region. One is directed. The goal of a directed audit is to verify a certain value such as a violation of a standard. The other is undirected. The goal of the undirected audit is to determine the quality of the data being generated. Both are systems audits but the undirected audit will have wider ramifications. For the most part, I will address the undirected audit.

Each measurement system requires a different auditing path. I point this out not to make the job sound complicated, but to emphasize the importance of having the auditor to be knowledgeable in the area of the audit. The path of auditing will be determined by:

- the quantity and quality of records,
- the existance of an agency SOP,
- the availability of records,
  - on a macro geographic scale,
  - on a micro geographic scale,
- the system itself,
- the time frame allowed.

Thus you can see that the auditing process is tailored to the specific system being audited.

In Region IV where every funded state and local agency is audited at least once a year this is the approach that we take.

1. Establish the flow of samples and data through the system (from the agency SOP).

2. Trace each parameter of the measurement process (volume, time, flowrate, etc.) back to the base standard and verify the quality of that standard.

3. Verify that all measurements and transfers of data are documented and follow reference methods.

4. Verify that all measurements, calculations, and data transfers are accurate.

5. Provide feedback to the agency being audited of improvements that could be made in the measurement process as well as the data handling.

One of the most important aspects of this audit is that the agency itself has to participate and will itself determine the best corrective action for their own system.

DATA VALIDATION FOR THE LOS ANGELES CATALYST STUDY (LACS)

by

Charles E. Rodes

Environmental Monitoring Systems Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

DATA VALIDATION FOR THE LOS ANGELES CATALYST STUDY (LACS)

C.E. Rodes

INTRODUCTION

The Environmental Monitoring and Support Laboratory (EMSL) is very concerned with the quality of data generated in its field studies. This is reflected in the quality control measures employed by EMSL during sampling and analysis, and the data validation performed before data are released.

Data validation like other aspects of quality control requires resource allocations, especially in terms of the manpower required to complete the final validation. The amount or degree of validation required is dependent upon the end use of the data. In a study such as the Los Angeles Catalyst Study (LACS) which is primarily concerned with long-term trends, the emphasis in data validation is to detect any extreme outliers which would affect monthly averages. Since we do not report maximums, our data validation philosophy for this study is primarily concerned with those values that may affect long-term averages.

CONCLUSIONS

As the project officer responsible for the study, I initially chose an acceptance error band of ±10% on individual measured values, hence, this is also the error band of the averages generated from these numbers.

Given this requirement one should be able to assess statistically the amount and types of validation required to prevent data reduction and transfer errors from contributing more than 1 to 2% to this overall ±10% error. Unfortunately this area has really not been examined for this study in any detail nor I expect for many other studies. The present validation levels used for the LACS probably examine more data than are necessary to maintain the desired error level, but in regard to validation, I would much rather be conservative than embarrassed after the data are released.

PROCEDURES

The main objective of the Los Angeles Catalyst Study (LACS) is to develop ambient air data bases for sulfate ($SO_4^=$), carbon monoxide (CO), lead (Pb), and other mobile source related pollutants before and after introduction of the 1975-model automobiles that employ catalytic converters. The data from this study are being analyzed to determine whether the catalytic converter has significantly increased the ambient sulfate levels and/or simultaneously decreased the ambient CO and Pb levels near the San Diego Freeway in Los Angeles.

The Environmental Monitoring and Support Laboratory (EMSL) is responsible for all study-related functions including instrumentation, operation, sample analyses, quality control, and data validation and analyses. Since January 1976, the operation of instruments and analyses of samples were performed under contract to Rockwell International or by interagency agreement with the Lawrence Berkeley Laboratory. To assure the quality of the data supplied by these two organizations, EMSL maintains a comprehensive quality assurance program covering all aspects of the study. EMSL issues periodic reports which discuss the trends and the interrelationships among the various pollutant patterns.

The site locations in Los Angeles and the site layouts in relation to the San Diego Freeway are shown in Figure 1. By selecting sites with the prevailing wind perpendicular to the freeway, the cross-freeway contribution to the ambient pollutant levels can be determined using concurrent upwind and downwind measurements.

The data collected are classed as either continuous or integrated depending on the measurement method. Continuous data are reduced to hourly averages and integrated data are collected either over a 4-hour or 24-hour period. The total data volume generated by the LACS is shown in Table 1. Since the sites are usually shut down in December of each year for routine maintenance, the data volumes are based on an 11-month year.

The flow of samples and data are shown in Figure 2. All block items except "Data Processing at RTP" and "Final Data Validation" are performed by the contractor. Data validation steps taken by the contractor are referred to as "pre-validation", while validation performed at RTP under more direct EPA control are referred to as "final validation".

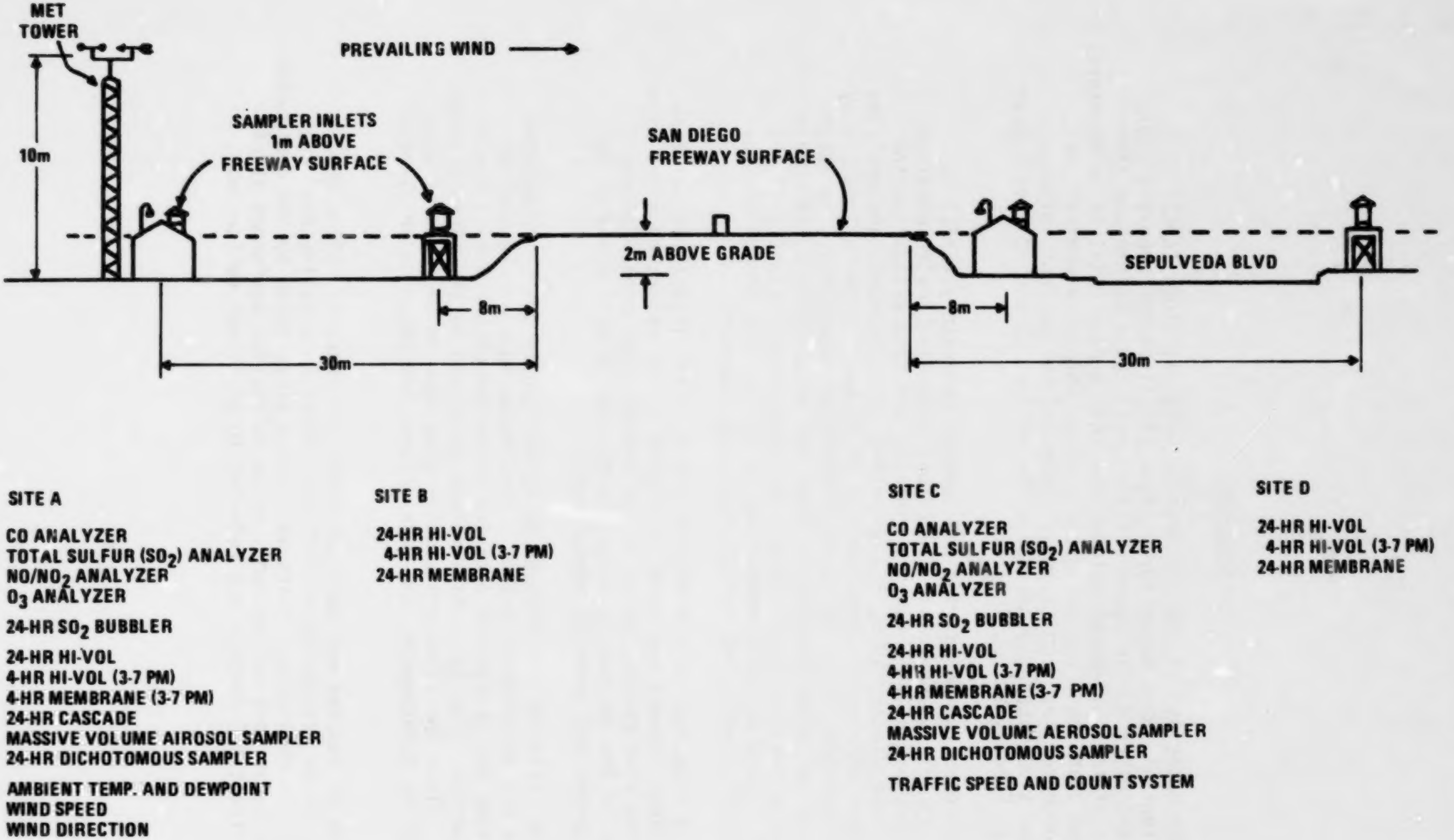

MET TOWER
10m
PREVAILING WIND →
SAMPLER INLETS 1m ABOVE FREEWAY SURFACE
SAN DIEGO FREEWAY SURFACE
2m ABOVE GRADE
SEPULVEDA BLVD
8m
30m
8m
30m
SITE A
CO ANALYZER
TOTAL SULFUR (SO_2) ANALYZER
NO/NO_2 ANALYZER
O_3 ANALYZER
24-HR SO_2 BUBBLER
24-HR HI-VOL
4-HR HI-VOL (3-7 PM)
4-HR MEMBRANE (3-7 PM)
24-HR CASCADE
MASSIVE VOLUME AIROSOL SAMPLER
24-HR DICHOTOMOUS SAMPLER
AMBIENT TEMP. AND DEWPOINT
WIND SPEED
WIND DIRECTION
SITE B
24-HR HI-VOL
4-HR HI-VOL (3-7 PM)
24-HR MEMBRANE
SITE C
CO ANALYZER
TOTAL SULFUR (SO_2) ANALYZER
NO/NO_2 ANALYZER
O_3 ANALYZER
24-HR SO_2 BUBBLER
24-HR HI-VOL
4-HR HI-VOL (3-7 PM)
4-HR MEMBRANE (3-7 PM)
24-HR CASCADE
MASSIVE VOLUME AEROSOL SAMPLER
24-HR DICHOTOMOUS SAMPLER
TRAFFIC SPEED AND COUNT SYSTEM
SITE D
24-HR HI-VOL
4-HR HI-VOL (3-7 PM)
24-HR MEMBRANE

Figure 1. LACS study site composition and elevation.

Table 1. LACS YEARLY* DATA VOLUME

	SUMMER	WINTER	TOTAL
CONTINUOUS (HOURLY)	70,080	58,560	128,640
INTEGRATED (4-HR)	20,160	11,100	31,260
INTEGRATED (24-HR)	9,000	4,050	13,050
INTEGRATED (WEEKLY)	816	680	1,496
INTEGRATED MONTHLY)	120	100	220
	100,176	74,490	174,666

*ASSUMES OPERATION FOR 11 MONTHS/YEAR.

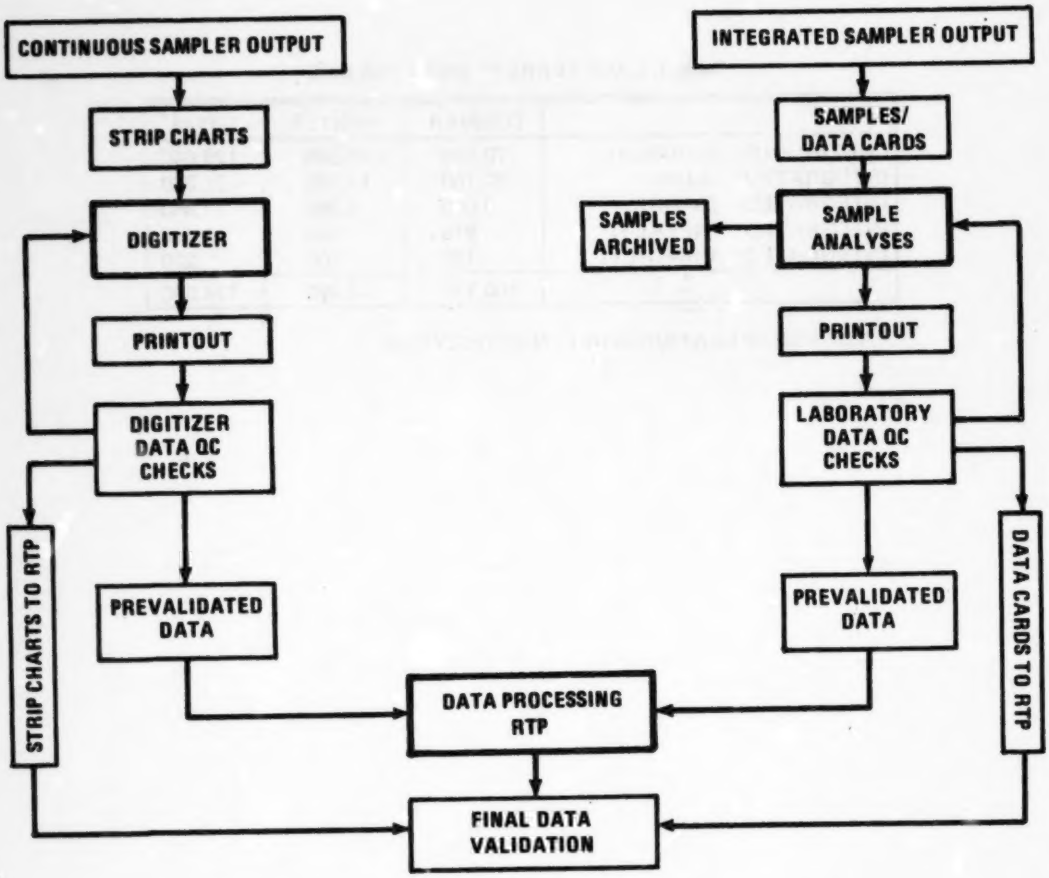

CONTINUOUS SAMPLER OUTPUT
STRIP CHARTS
DIGITIZER
PRINTOUT
DIGITIZER DATA QC CHECKS
PREVALIDATED DATA
STRIP CHARTS TO RTP
INTEGRATED SAMPLER OUTPUT
SAMPLES/ DATA CARDS
SAMPLES ARCHIVED
SAMPLE ANALYSES
PRINTOUT
LABORATORY DATA QC CHECKS
PREVALIDATED DATA
DATA CARDS TO RTP
DATA PROCESSING RTP
FINAL DATA VALIDATION

Figure 2. LACS Data Flow.

Pre-validation by the contractor is performed in two areas - electronic digitization of the strip charts and compilation of the analysis data in the laboratory. A portion of the data generated by the electronic digitizer are checked against manually read strip charts to verify scaling and digitizer performance. At present 5% of the data are spot checked in this procedure. The laboratory analysis results are compared on the contractors computer listing against the data cards manually completed during the analyses. All data (100%) generated in the laboratory are checked in this procedure because of the importance of single integrated values. We do not at present require the contractor to keep records of the amount of data corrected during prevalidation.

Final data validation is performed at RTP following the general procedure in Figure 3. This step in the validation is concerned primarily with data transfer errors, but also examines data that are not consistent (outliers) with the rest of the data base. In general all of the values in approximately the highest and lowest 1.0 percentile are verified with a check made at random of approximately 5% of the remaining data. These validation levels were initially selected somewhat arbitrarily by the project officer as a compromise between data quality and the amount of resources required for the validation.

The final data validation procedures are based upon the output formats used to list the individual data values. The three formats are: (1) an hourly listing for continuous data such as CO and NO, (2) an integrated data listing for samples averaged over 4-hour or 24-hour periods, and (3) a summary listing comparing simultaneously collected upwind and downwind data for freeway contribution. The general instructions given to the data clerks are shown in Figure 4. A sample printout of hourly data is shown in Figure 5, followed by the outlier limits in Table 2 used in validating the hourly data. A sample of a 24-hour integrated data printout is shown in Figure 6 with its associated validation limits given in Table 3. A study is presently being made of the frequency distributions of the LACS data to reassess the validation limits listed in Tables 2 and 3. The starred (*) values on the printouts are values determined to be outside ±3 standard deviations of the monthly means. A sample of the summary format is shown in Figure 7.

For possible future validation requirements, portions of integrated samples are stored at the contractor's laboratory, and the strip charts, data cards, and final validated printouts are stored by EPA at RTP.

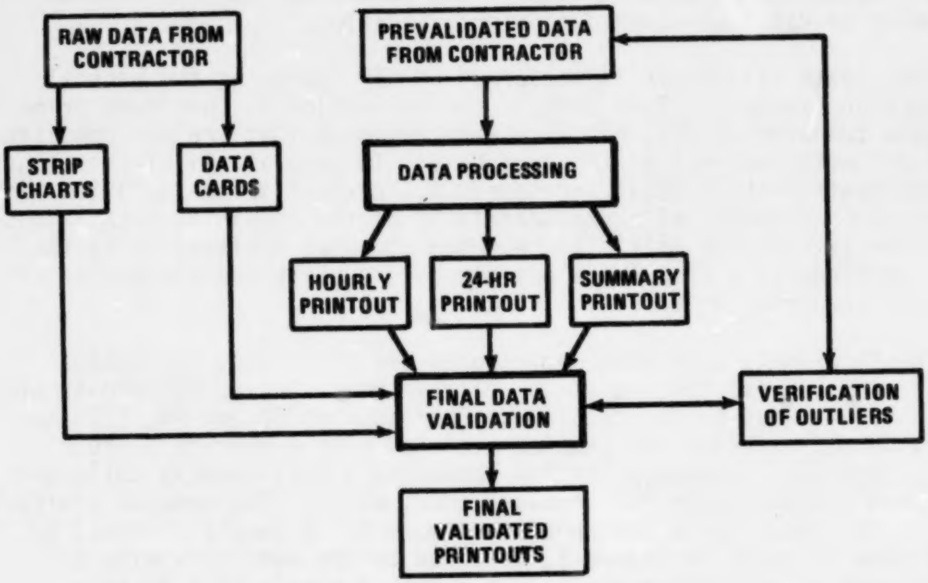

RAW DATA FROM CONTRACTOR
PREVALIDATED DATA FROM CONTRACTOR
STRIP CHARTS
DATA CARDS
DATA PROCESSING
HOURLY PRINTOUT
24-HR PRINTOUT
SUMMARY PRINTOUT
FINAL DATA VALIDATION
VERIFICATION OF OUTLIERS
FINAL VALIDATED PRINTOUTS

Figure 3. Final data validation.

Figure 4. LACS PRINTOUT VALIDATION INSTRUCTIONS

GENERAL

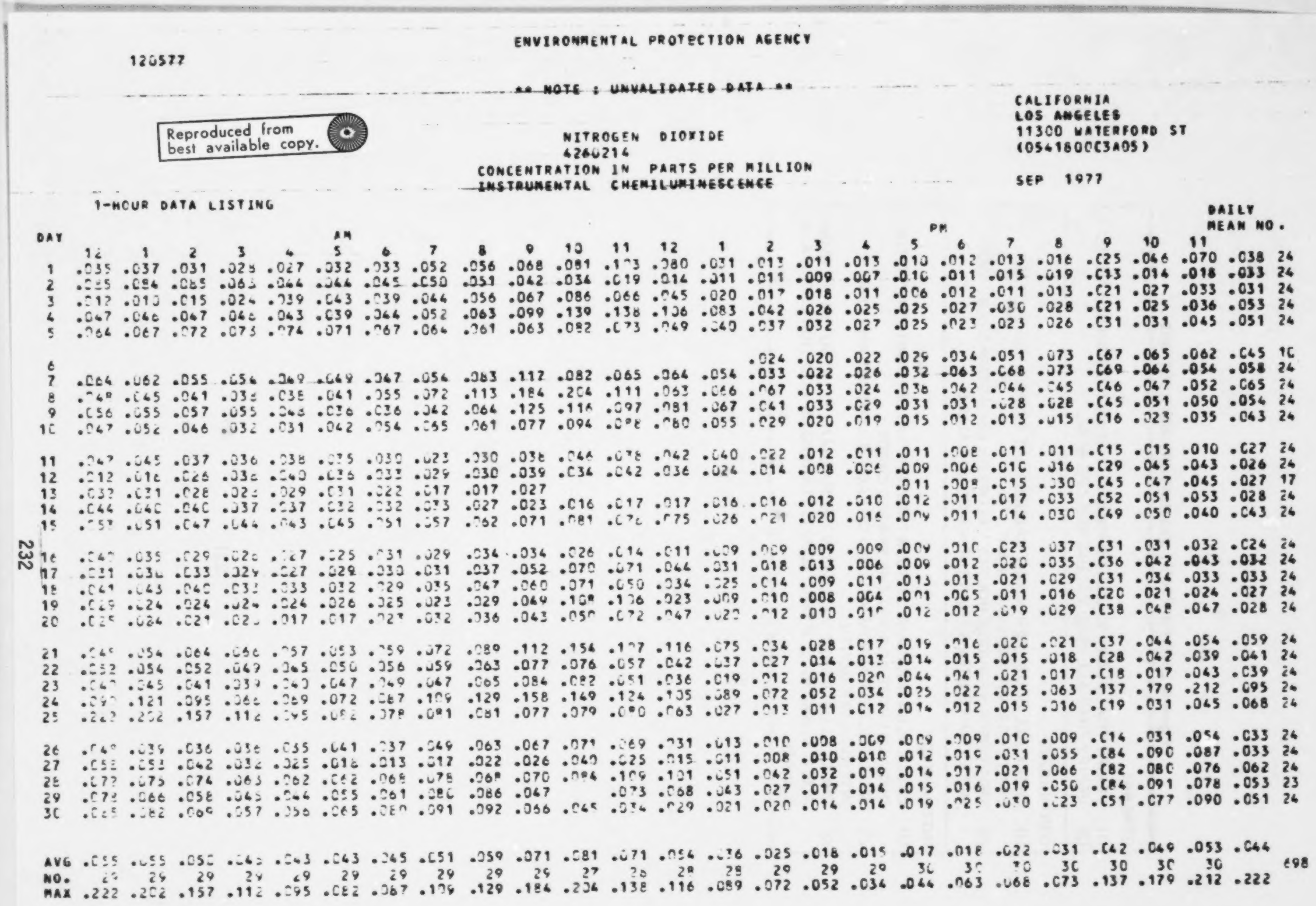
(1) VERIFY THAT BLANK SPACES ON PRINTOUT MEAN THAT NO DATA EXISTS.
(2) VERIFY THAT ALL ZERO VALUES (0.0) ARE REAL.

CONTINUOUS

(1) CHECK ALL HOURLY PRINTOUT VALUES THAT EXCEED THE OUTLIER LIMITS AGAINST THE STRIP CHART.
(2) SPOT CHECK 5 RANDOM HOURLY VALUES ON EACH STRIP CHART (ONE WEEK/CHART) OTHER THAN THE MAXIMUM VALUES.

INTEGRATED

(1) CHECK ALL 4-HOUR AND 24-HOUR PRINTOUT VALUES THAT EXCEED THE OUTLIER LIMITS AGAINST THE SAROAD CARDS.
(2) SPOT CHECK 2 RANDOM VALUES FOR EACH POLLUTANT AND TIME INTERVAL PER MONTH.
(3) CHECK ALL STARRED VALUES ON THE SUMMARY PRINTOUT IN COLUMNS A, B, C, D, AND (C-A). IF ONLY (C-A) IS STARRED, IN ADDITION CHECK A, B, C, and D.



```
                                   ENVIRONMENTAL PROTECTION AGENCY
      120577
                                    ** NOTE : UNVALIDATED DATA **
                                                                                  CALIFORNIA
                                                                                  LOS ANGELES
    Reproduced from                       NITROGEN   DIOXIDE                      11300 WATERFORD ST
    best available copy.                  4260214                                 (054180003A05)
                                   CONCENTRATION IN  PARTS PER MILLION
                                   INSTRUMENTAL  CHEMILUMINESCENCE                SEP  1977
   1-HOUR DATA LISTING
                                                                                                  DAILY
                                                                        PM                        MEAN NO.
DAY                          AM
     12    1    2    3    4    5    6    7    8    9   10   11   12    1    2    3    4    5    6    7    8    9   10   11
 1 .035 .037 .031 .028 .027 .032 .033 .052 .056 .068 .081 .103 .080 .031 .013 .011 .013 .010 .012 .013 .016 .025 .046 .070 .038 24
 2 .025 .084 .065 .063 .044 .044 .045 .050 .051 .042 .034 .019 .014 .011 .011 .009 .007 .010 .011 .015 .019 .013 .014 .018 .033 24
 3 .012 .010 .015 .024 .039 .043 .039 .044 .056 .067 .086 .066 .045 .020 .017 .018 .011 .006 .012 .011 .013 .021 .027 .033 .031 24
 4 .047 .046 .047 .046 .043 .039 .044 .052 .063 .099 .139 .138 .106 .083 .042 .026 .025 .025 .027 .030 .028 .021 .025 .036 .053 24
 5 .064 .067 .072 .073 .074 .071 .067 .064 .061 .063 .082 .073 .049 .040 .037 .032 .027 .025 .023 .023 .026 .031 .031 .045 .051 24

 6                                                                  .024 .020 .022 .029 .034 .051 .073 .067 .065 .062 .045 10
 7 .064 .062 .055 .054 .049 .049 .047 .054 .083 .112 .082 .065 .064 .054 .033 .022 .026 .032 .063 .068 .073 .069 .064 .054 .058 24
 8 .048 .045 .041 .038 .038 .041 .055 .072 .113 .184 .204 .111 .063 .066 .067 .033 .024 .036 .042 .044 .045 .046 .047 .052 .065 24
 9 .056 .055 .057 .055 .048 .036 .036 .042 .064 .125 .116 .097 .081 .067 .041 .033 .029 .031 .031 .028 .028 .045 .051 .050 .054 24
10 .047 .052 .046 .032 .031 .042 .054 .065 .061 .077 .094 .088 .080 .055 .029 .020 .019 .015 .012 .013 .015 .016 .023 .035 .043 24

11 .047 .045 .037 .036 .038 .035 .030 .023 .030 .038 .046 .078 .042 .040 .022 .012 .011 .011 .008 .011 .011 .015 .015 .010 .027 24
12 .012 .016 .026 .038 .040 .036 .033 .029 .030 .039 .034 .042 .036 .024 .014 .008 .006 .009 .006 .010 .016 .029 .045 .043 .026 24
13 .032 .031 .028 .024 .029 .031 .022 .017 .017 .027                               .011 .009 .015 .030 .045 .047 .045 .027 17
14 .044 .040 .040 .037 .037 .032 .032 .033 .027 .023 .016 .017 .017 .016 .016 .012 .010 .012 .011 .017 .033 .052 .051 .053 .028 24
15 .053 .051 .047 .044 .043 .045 .051 .057 .062 .071 .081 .076 .075 .026 .021 .020 .016 .009 .011 .014 .030 .049 .050 .040 .043 24

16 .040 .035 .029 .028 .027 .025 .031 .029 .034 .034 .026 .014 .011 .009 .009 .009 .009 .009 .010 .023 .037 .031 .031 .032 .024 24
17 .031 .036 .033 .029 .027 .029 .030 .031 .037 .052 .070 .071 .044 .031 .018 .013 .006 .009 .012 .020 .035 .036 .042 .043 .032 24
18 .041 .043 .040 .032 .033 .032 .029 .035 .047 .060 .071 .050 .034 .025 .014 .009 .011 .013 .013 .021 .029 .031 .034 .033 .033 24
19 .029 .024 .024 .024 .024 .026 .025 .023 .029 .049 .108 .106 .023 .009 .010 .008 .004 .001 .005 .011 .016 .020 .021 .024 .027 24
20 .025 .024 .021 .020 .017 .017 .023 .032 .036 .043 .050 .072 .047 .020 .012 .010 .010 .012 .012 .019 .029 .038 .048 .047 .028 24

21 .049 .054 .064 .066 .057 .053 .059 .072 .089 .112 .154 .107 .116 .075 .034 .028 .017 .019 .016 .020 .021 .037 .044 .054 .059 24
22 .052 .054 .052 .049 .045 .050 .056 .059 .063 .077 .076 .057 .042 .037 .027 .014 .013 .014 .015 .015 .018 .028 .042 .039 .041 24
23 .040 .045 .041 .039 .040 .047 .049 .067 .065 .084 .082 .051 .036 .019 .012 .016 .020 .044 .041 .021 .017 .018 .017 .043 .039 24
24 .090 .121 .095 .066 .069 .072 .087 .109 .129 .158 .149 .124 .105 .089 .072 .052 .034 .025 .022 .025 .063 .137 .179 .212 .095 24
25 .222 .202 .157 .112 .095 .082 .078 .081 .081 .077 .079 .080 .063 .027 .013 .011 .012 .014 .012 .015 .016 .019 .031 .045 .068 24

26 .049 .039 .036 .038 .035 .041 .037 .049 .063 .067 .071 .069 .031 .013 .010 .008 .009 .009 .009 .010 .009 .014 .031 .054 .033 24
27 .058 .053 .042 .032 .025 .018 .013 .017 .022 .026 .040 .025 .015 .011 .008 .010 .010 .012 .019 .031 .055 .084 .090 .087 .033 24
28 .072 .075 .074 .063 .062 .062 .068 .078 .068 .070 .084 .109 .101 .051 .042 .032 .019 .014 .017 .021 .066 .082 .080 .076 .062 24
29 .073 .066 .058 .045 .044 .055 .061 .080 .086 .047      .073 .068 .043 .027 .017 .014 .015 .016 .019 .050 .084 .091 .078 .053 23
30 .065 .062 .069 .057 .056 .065 .060 .091 .092 .066 .045 .034 .029 .021 .020 .014 .014 .019 .025 .030 .023 .051 .077 .090 .051 24


AVG .055 .055 .050 .045 .043 .043 .045 .051 .059 .071 .081 .071 .054 .036 .025 .018 .015 .017 .018 .022 .031 .042 .049 .053 .044
NO.   29   29   29   29   29   29   29   29   29   29   27   28   28   28   29   29   29   30   30   30   30   30   30   30        698
MAX .222 .202 .157 .112 .095 .082 .087 .109 .129 .184 .204 .138 .116 .089 .072 .052 .034 .044 .063 .068 .073 .137 .179 .212 .222
```

Figure 5. Sample hourly printout.

Table 2. LACS CONTINUOUS SAMPLER OUTLIER LIMITS

LACS CONTINUOUS SAMPLER OUTLIER LIMITS (ppm)		
	SITE 008	ALL OTHERS
CO (CARBON MONOXIDE)	25.0	15.0
NO (NITRIC OXIDE)	–	0.5
NO_2 (NITROGEN NIOXIDE)	–	0.3
O_3 (OZONE)	–	0.3
TS (TOTAL SULFUR)	–	0.05
WS (WIND SPEED)	–	15[A]

[A] MILES/HOUR

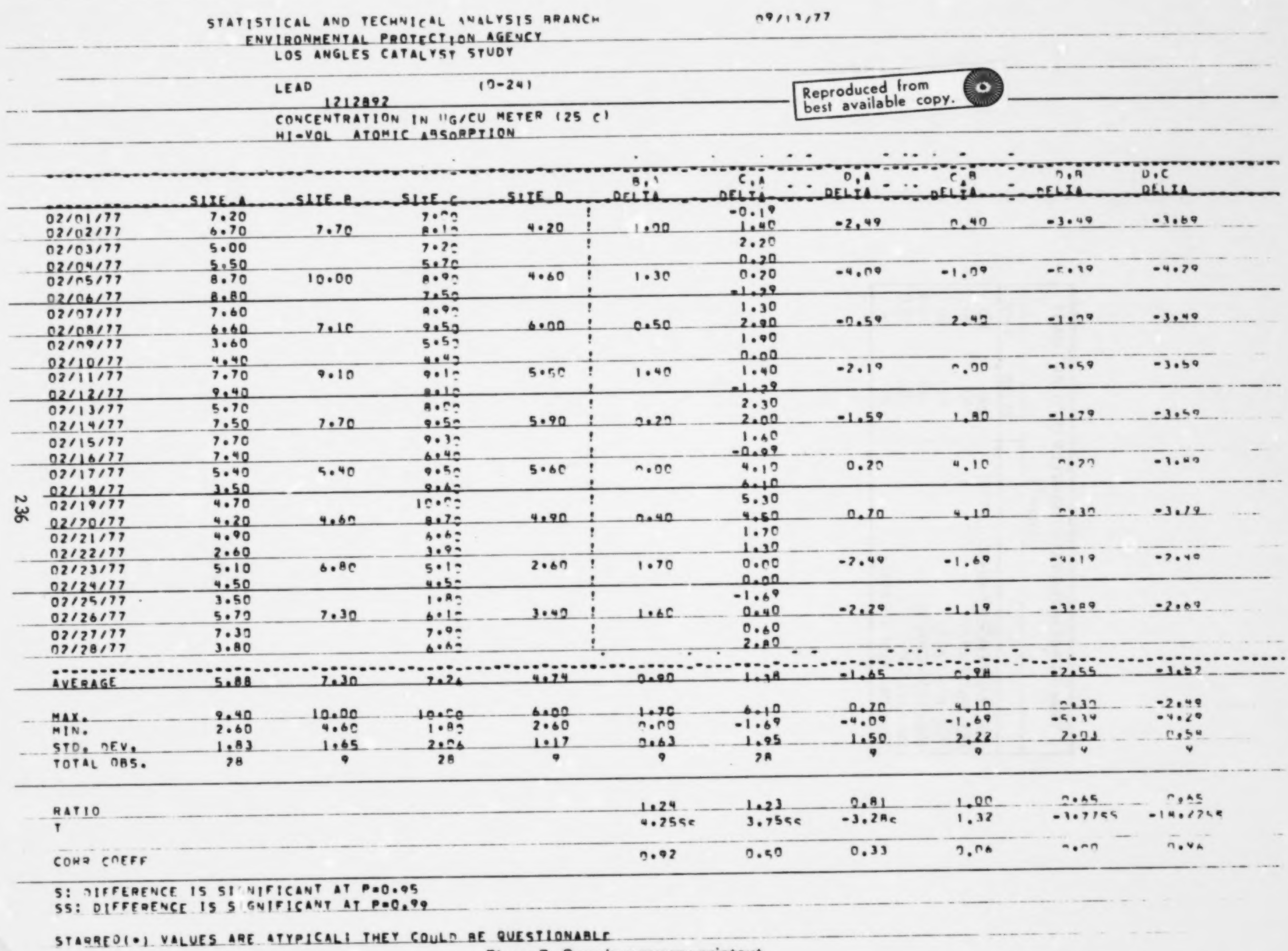
Reproduced from best available copy.

```
                                                 STATISTICAL SERVICES STAFF                    10/13/77
                                               ENVIRONMENTAL PROTECTION AGENCY
                                                     24-HOUR DATA LISTING

                                                                                  054180005A05
                                                                                  CALIFORNIA
                                                                                  LOS ANGELES
                                                                                  SITE 005

              SUSPENDED    NITRATE      SULFATE      LEAD         AMMONIUM     SULFUR
              PART.                                                            DIOXIDE
  YR/MO/DAY   1110191017   1230692017   1240391017   1212892017   1230192017   4240191017

  77/01/24        60          3.34          4.0         3.40         0.44
  77/01/26       124          6.42          4.5         5.10         0.59
  77/01/27       102         13.01          6.2         5.50         2.43          37
  77/01/28       112         12.73          6.8         8.40         3.20          27
  77/01/29        81          8.04          5.3         7.60         1.11
  77/01/30        96         13.89          6.8         9.70         2.60          37
  77/01/31       124         17.87          8.4         8.40         4.24          41

  77/02/01       124         18.82          9.1         7.00         4.52          46
  77/02/02       121         16.79          7.0         8.10         3.54          41
  77/02/03        82          6.27          3.4         7.90         0.79          34
  77/02/04        76          2.93          2.7         5.70         0.50          25
  77/02/05        84          8.90          4.4         8.00         1.43          42
  77/02/06       132         24.67          8.9         7.50         6.20*         60
  77/02/07       132         15.48          7.6         8.00         2.75          42
  77/02/08       157         21.05         10.3         9.50         5.61          27
  77/02/09       105          7.24          4.4         5.50         1.12          17
  77/02/10        73          2.92          4.8         4.40         0.73          50
  77/02/11       113          9.45          4.4         9.10         1.34          40
  77/02/12        81          2.09          4.5         8.10         0.97          32
  77/02/13        81          7.86          4.0         8.00         0.89
  77/02/14       133         13.80          9.5         9.50         1.43          39
  77/02/15        93          5.10          3.1         2.30         0.40          33
  77/02/16        95          5.07          4.0         6.40         0.29          13
  77/02/17       133         19.03          9.1         9.50         2.49          59
  77/02/18       150          9.41         21.3         9.60         3.25          51
  77/02/19       124         10.95         19.4        10.00         4.36          37
  77/02/20        94         10.54         14.7         8.70         1.93
  77/02/21        95         10.43          9.4         6.60         0.93
  77/02/22        84          2.05          4.9         3.00         0.09*         20
  77/02/23        65          4.98          6.0         5.10         0.26          29
  77/02/24        57          2.92          4.7         4.00         0.21*         16
  77/02/25        48          3.65          2.5         1.90         3.09*
  77/02/26        45          7.74          3.6         6.10         0.36          26
  77/02/27        70          7.27          3.5         7.90         0.45
  77/02/28       102          7.23          4.0         6.60         0.30

  77/03/01       114          2.58          3.6         2.90         0.17*         17
```

Figure 6. Sample 24-hour printout.

Table 3. LACS INTEGRATED SAMPLER OUTLIER LIMITS ($\mu g/m^3$)

	24 HR	4 HR
TSP (SUSPENDED PARTICULATES)	200	300
NO_3^- (NITRATE)	30	30
$SO_4^=$ (SULFATE)	30	50
NH_4^+ (AMMONIUM)	3.0	3.0
Pb (LEAD)	8.0	12.0
SO_2 (SULFUR DIOXIDE)	50	–



Reproduced from best available copy.

```
                STATISTICAL AND TECHNICAL ANALYSIS BRANCH                 09/13/77
                  ENVIRONMENTAL PROTECTION AGENCY
                    LOS ANGLES CATALYST STUDY

                    LEAD                   (0-24)
                       1212892
                    CONCENTRATION IN UG/CU METER (25 C)
                    HI-VOL  ATOMIC ABSORPTION

-----------------------------------------------------------------------------------------------------------------
                                                          B,A       C,A       D,A       C,B       D,B       D,C
               SITE A    SITE B    SITE C    SITE D      DELTA     DELTA     DELTA     DELTA     DELTA     DELTA
02/01/77        7.20                7.00            !              -0.19
02/02/77        6.70      7.70      8.10      4.20  !     1.00      1.40     -2.49      0.40     -3.49     -3.89
02/03/77        5.00                7.20            !               2.20
02/04/77        5.50                5.70            !               0.20
02/05/77        8.70     10.00      8.90      4.60  !     1.30      0.20     -4.09     -1.09     -5.39     -4.29
02/06/77        8.80                7.50            !              -1.29
02/07/77        7.60                8.90            !               1.30
02/08/77        6.60      7.10      9.50      6.00  !     0.50      2.90     -0.59      2.40     -1.09     -3.49
02/09/77        3.60                5.50            !               1.90
02/10/77        4.40                4.40            !               0.00
02/11/77        7.70      9.10      9.10      5.50  !     1.40      1.40     -2.19      0.00     -3.59     -3.59
02/12/77        9.40                8.10            !              -1.29
02/13/77        5.70                8.00            !               2.30
02/14/77        7.50      7.70      9.50      5.90  !     0.20      2.00     -1.59      1.80     -1.79     -3.59
02/15/77        7.70                9.30            !               1.60
02/16/77        7.40                6.40            !              -0.99
02/17/77        5.40      5.40      9.50      5.60  !     0.00      4.10      0.20      4.10      0.20     -3.89
02/18/77        3.50                9.60            !               6.10
02/19/77        4.70               10.00            !               5.30
02/20/77        4.20      4.60      8.70      4.90  !     0.40      4.50      0.70      4.10      0.30     -3.79
02/21/77        4.90                6.60            !               1.70
02/22/77        2.60                3.90            !               1.30
02/23/77        5.10      6.80      5.10      2.60  !     1.70      0.00     -2.49     -1.69     -4.19     -2.49
02/24/77        4.50                4.50            !               0.00
02/25/77        3.50                1.80            !              -1.69
02/26/77        5.70      7.30      6.10      3.40  !     1.60      0.40     -2.29     -1.19     -3.89     -2.69
02/27/77        7.30                7.90            !               0.60
02/28/77        3.80                6.60            !               2.80
-----------------------------------------------------------------------------------------------------------------
AVERAGE         5.88      7.30      7.24      4.74        0.90      1.38     -1.65      0.98     -2.55     -3.52

MAX.            9.40     10.00     10.00      6.00        1.70      6.10      0.70      4.10      0.30     -2.49
MIN.            2.60      4.60      1.80      2.60        0.00     -1.69     -4.09     -1.69     -5.39     -4.29
STD. DEV.       1.83      1.65      2.06      1.17        0.63      1.95      1.50      2.22      2.01      0.59
TOTAL OBS.       28         9        28         9           9        28         9         9         9         9

RATIO                                                     1.24      1.23      0.81      1.00      0.65      0.65
T                                                         4.25SS    3.75SS   -3.28S     1.32     -3.77SS  -18.22SS

CORR COEFF                                                0.92      0.50      0.33      0.06      0.00      0.96

S: DIFFERENCE IS SIGNIFICANT AT P=0.95
SS: DIFFERENCE IS SIGNIFICANT AT P=0.99

STARRED(*) VALUES ARE ATYPICAL: THEY COULD BE QUESTIONABLE
```

Figure 7. Sample summary printout.

VALIDATION TECHNIQUES USED IN CONTINUOUS AIR MONITORING

by

Marvin B. Hertz

Health Effects Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

VALIDATION TECHNIQUES USED IN CONTINUOUS AIR MONITORING

M.B. Hertz

The Community Health Air Monitoring Program (CHAMP) is a network of air monitoring stations used to acquire reliable air quality data for use in epidemiologic health effects studies.

The CHAMP network has remote air monitoring stations located in each of the selected health study communities across the country. The focal point of the CHAMP network is the central computer facility located at the Environmental Protection Agency, Research Triangle Park, North Carolina. A minicomputer at each of the remote stations controls and acquires data and associated system status information from aerometric and meteorologic instrumentation and transmits the data by phone lines to the central computer facility. The central controller for the CHAMP network is a dual processor system with a full complement of input, storage, and display peripherals. One minicomputer was selected to perform the tasks associated with the management of the large data base to be generated by the network. The telecommunications and real-time processing tasks are handled by the other processor.

Two fundamental system objectives were: (1) to provide machine validation of the data, and (2) to develop a management information system for use in the quality assurance, field logistics, and field maintenance tasks associated with system operations.

Remote Data Acquisition System

Basically, the minicomputer in the remote station serves as an interface between the pollutant analyzers and associated system, magnetic tape data storage, the remote field service operator, and the telecommunications network. The data generated and recorded at the remotes and transmitted to central includes not only the actual meteorologic and pollutant sensor responses, but also associated analog signals and digital status signals.

These signals supply information about the performance and status of each instrument. For example, if an instrument is switched from an ambient sampling mode to the calibration mode, a status bit is recorded which reflects this change.

Telecommunication

Data is retrieved on the request of the central computer system from each of the remote stations via a dial up phone line at two-hour intervals. The central and remote computers converse via voice-grade telecommunications system consisting of modems operating in a full duplex mode at the rate of 1200 baud from remote to central and 150 baud in the reverse direction. Polling is under the complete control of the central controller. A file is maintained on a disk at central which contains the phone number of each station in the format required by the calling software system. An alterable polling queue is also disk resident. A rigid protocol has been established to guarantee accurate transmission and retrieval of data. Central makes several tries to establish contact with a remote station before abandoning the attempt and placing the station at the bottom of the polling queue. A hardware carrier detect protocol establishes the communications link. Each frame is checked for parity and framing errors by the modem controller. Checksums are computed for each 512 frame record and compared by the computer. An acknowledge character is exchanged indicating correct receipt of the record. Should any of the tests fail, several transmission retries are made. Communications are terminated by receipt of a character from the remote system indicating the end of data or by failure of the remote to transmit in the required period of time.

The Central Controller Hardware Configuration

The focal point of the CHAMP network is the central computer facility located at the Environmental Protection Agency Complex, Research Triangle Park, North Carolina. The central controller for the CHAMP network is a dual processor system with a full complement of input, storage, and display peripherals. The heavy burden on processor time placed by the telecommunications and real-time processing of the large quantities of data justified the choice of a dual processor system. A PDP-11/40 with 40K of core was selected to perform the tasks associated with the management of the large data base generated by the network. The telecommunications and real time processing tasks are handled by a PDP-11/05 computer with 16K of core. The two processors are interconnected by a Unibus window which takes advantage of the unified asynchronous data path architecture of the 11 system. The window allows each processor to address the core and peripherals on the other processor as if it were its own. In addition, the DEC memory management option was added to the PDP-11/40 to handle addressing above 32K in the 16-bit system. An extensive complement of peripherals including two 1.2M word cartridge type disks, three tape drives, and electrostatic printer-plotter, line printer, and CRT display were initially selected; the rapid retrieval requirement for large quantities of data necessitated the addition of a Telefile dual spindle, quad density, removable 20 surface pack disk system capable of storing 98M words.

The Central Controller Software Requirements

As mentioned previously, the PDP-11/05 processor is dedicated to the system telecommunication tasks and the storing of the data simultaneously on a magnetic tape and the Telefile disk as received. The data (Level 1) so stored is an image of the tapes recorded at the remote station. These data include the primary data (those data which actually represent parameters of interest such as pollution levels), secondary data (that data required to validate primary data or which are used only to insure proper station operation), and the status bits. For flexibility any channel at the remote stations can be selectively assigned a primary or secondary function as required; furthermore, the number of primary and secondary channels is made arbitrary. All of the data at the remote stations, whether primary or secondary data, is assigned a remote station data slot (RSDS). The complement of instruments and the RSDS number corresponding to a given instrument may be different in each station. It is, therefore, necessary to append a "map" which gives the correspondence between instrument and RSDS number at the front of each set of station data. At the central each parameter is assigned a mnemonic (2 to 4 letters) which describes the parameter (NOS, O3, TOUT, etc.). The map, therefore, must contain the mnemonic and the link between the mnemonic and the corresponding RSDS. As the complement of instruments in a station changes, a new map will be created and appended to future data from that station. All operations at Central, will, therefore, refer only to the mnemonic names and not any "channel," "data slot," or other number.

The 11/40 processor is devoted to data validation tasks and task associated with the management information system requirements for quality assurance, field maintenance, and field logistics.

CHAMP Software Features

As mentioned previously, fundamental system objectives were:

1. To develop and implement a computer-based management information system for use in system quality assurance, field logistics, and field maintenance tasks.
2. To provide for machine (computer)-validation of the data.

The current CHAMP software, which will now be described, represents the composite of original system programs plus those that were subsequently developed in response to needs recognized after the system became operational and to inadequacies in the original software.

Current CHAMP Central Software

1. FILMAP - This program creates the air quality data base station map files. The station map files contain station location information, instrument complement by station, status bit configuration indicating hardware failure, calibration data, and correct operating ranges of the hardware. In addition, the map files identify validation criteria such as secondary parameter limits, dependencies between primary parameters, filtering and interpolation techniques to be applied, and the format of the final data output.

2. FILSET - This program sorts polled and station data (mailed) by type and writes the data in the correct format into the appropriate files in the data base. Six types of data are sorted. These are (1) primary parameter data, (2) calibration constants, (3) status words, (4) journal entries, (5) secondary parameter data, and (6) calibration data. Table 1 presents a sample of secondary data.

3. TIMSTN - This program processes the remote station data tapes and checks for remote station data time anomalies. The program enables

the operator to edit the reported times to resolve the time jumps. The edited station data is recorded on magnetic tape.

4. PDAILY - This program summarizes and produces a printed report of daily station performance. PDAILY invalidates primary parameter data in cases where bits associated with hardward failures are set. Calibration data and data which is collected on the wrong instrument range is invalidated. The number of invalid five-minute averages for each primary parameter for each hour is tallied as well as the number of times each status bit is set for each hour. The amount of data found invalid, valid, missing or in calibration made for each primary parameter is summarized over the day. The journal entries, which are operator comments entered at the remote stations, are also listed.

5. PSUMRY - This program summarizes daily station performance and produces printed summaries of station performance. The performance summary contains the following:

    1. Percentage of valid data by parameter by day.
    2. Percentage of valid, missing, and calibration data by day.
    3. Percentage of valid, missing, and calibration data over the days shown on the summary.
    4. Logs of data processing progress.
    5. Primary and secondary parameter calibration occurrences, and occurrence of control chart samples which exceed upper or lower control chart limits.

6. PSCHRT - This program samples the air quality data base secondary parameters and generates control chart files from the sampled secondary parameter file data.

7. REVMAP - This program allows manual editing of the data by processing validation actions entered on punched cards in the standard "Review Change Request" format.

8. PCHEMS - This program generates the calibration, performs pre-established validation tests, and produces a printed report summarizing chemical analysis data (i.e., hi vol data, bubbler data).

9. PPCHRT - This program samples the air quality data base primary calibration parameters (A and B constants) and generates control chart files for these data.

10. PCALIB - This program generates the calibration coefficients used to convert the air quality data collected as a voltage by the remote station into a concentration. The calibration constants are calculated from the raw calibration mode data recorded at the remote station on tape.

11. FAUT - This program performs automatic calibration for selected equipment under control of the Central Computer and/or the remote station operator.

12. FILCAN - This program sorts the chemical analysis data by station by data and creates chemical analysis data files.

13. PLOTCC - This program takes input from the control chart files and plots control charts for the primary parameter zeroes, spans and B coefficients and for the secondary parameter ranges and values.

14. FLMRG - This program sorts station data (mailed) by types and writes the data in the correct format into the appropriate file in the data base, then these files are compared with those created by the FILSET Program (polled data) and fills the gaps in the data base.

TABLE 1. FIRST DATA WINDOW OF NORMALIZED SECONDARY PARAMETER STATISTICS AT STATION 3010 FOR JULIAN DAYS 77222 TO 77237

Sampling Period				Flow Rate, cc/min								
Beginning		Ending		Minimum			Maximum			Mean	Standard	Number of
Date	Time	Date	Time	Value	Date	Time	Value	Date	Time	Value	Deviation	Values
Nitrous Oxide												
222	1433	226	1042	124.80	225	1207	150.25	223	0906	146.81	2.6518	719
226	1241	228	1213	140.67	227	1508	144.50	228	0941	142.79	0.91668	527
228	1424	230	1145	142.31	229	0411	145.60	230	0852	143.58	0.81650	502
230	1222	232	1024	141.77	231	1528	147.92	232	0904	143.79	0.86354	524
232	1230	234	0911	143.00	232	1449	146.42	234	0906	144.69	0.47975	505
234	1114	236	0953	143.27	235	1637	145.87	235	0912	144.69	0.66924	519
Oxygen												
222	1433	226	1042	26.209	225	1822	31.173	226	0502	30.543	0.48780	727
226	1241	228	1213	30.092	227	1334	31.148	227	0850	30.631	0.22761	527
228	1424	230	1145	29.772	230	1039	30.509	229	2309	30.258	0.16089	502
230	1222	232	1024	25.644	232	0859	30.460	230	1928	30.004	0.47733	530
232	1230	234	0911	29.649	233	1114	31.124	234	0906	30.480	0.27402	505
234	1114	236	0953	29.600	235	1457	31.198	234	1119	30.298	0.27185	519
Ozone												
222	1514	226	1217	1019.6	224	0529	1062.7	226	1042	1040.0	10.286	699
226	1302	228	1412	943.43	228	0809	1068.1	227	0918	1021.2	31.643	539
228	1520	230	1207	921.88	228	2148	1045.0	229	1106	1031.4	13.212	377
230	1311	232	1251	1008.1	231	0920	1042.7	230	1903	1031.8	6.9196	537
234	1150	236	1106	1031.9	235	1642	1041.2	235	2148	1037.1	1.9239	537
232	1324	234	1129	1030.4	233	1119	1042.7	233	1419	1038.7	0.63916E-32	512

TABLE 1. (Continued)

Ethylene

Sampling Period				Flow Rate, cc/min								
Beginning		Ending		Minimum			Maximum			Mean	Standard	Number of
Date	Time	Date	Time	Value	Date	Time	Value	Date	Time	Value	Deviation	Values
222	1514	226	1217	17.247	225	1752	28.525	223	1918	24.976	0.75691	727
226	1302	228	1412	24.470	227	0506	25.665	228	0926	24.967	0.27847	539
228	1520	230	1207	25.019	230	1049	26.279	229	0855	25.546	0.21564	501
230	1311	232	1251	24.825	231	1523	25.827	231	0831	25.390	0.20101	537
234	1150	236	1106	25.003	234	1554	25.778	235	0912	25.358	0.15962	537
232	1324	234	1129	24.744	233	1114	25.827	234	1050	25.357	0.16027	512

Vacuum for Nitrous Oxide and Oxygen Samples

Sampling Period				Absolute Pressure, torr								
Beginning		Ending		Minimum			Maximum			Mean	Standard	Number of
Date	Time	Date	Time	Value	Date	Time	Value	Date	Time	Value	Deviation	Values
222	1433	226	1042	121.76	222	1843	132.45	224	0215	125.50	2.3853	649
226	1241	228	1213	122.53	226	1559	127.42	227	0254	126.02	1.2997	527
228	1424	230	1145	123.69	229	0526	126.39	228	2216	125.09	0.69572	502
230	1222	232	1024	114.72	232	0904	136.59	232	0739	122.56	4.2046	524
232	1230	234	0911	122.40	234	0832	127.94	234	0847	123.44	1.0806	505
234	1114	236	0953	117.92	235	1219	135.04	236	0714	126.88	3.6658	519

Vacuum for Ozone and Ethylene Samples

222	1512	226	1236	121.76	222	1843	132.45	224	0215	125.51	2.3831	646
226	1301	228	1459	122.53	226	1559	127.42	227	0254	126.03	1.2845	535
228	1520	230	1247	123.17	230	1247	126.39	228	2216	125.04	0.71975	500
230	1312	232	1251	114.72	232	0904	136.59	232	0739	122.57	4.2224	520
234	1150	236	1131	117.92	235	1219	135.04	236	0714	126.75	3.7078	532
232	1324	234	1129	122.40	234	1129	127.94	234	0847	123.43	1.0771	504

TABLE 1. (Continued)

Barometric Pressure For All Samples In Table 1

Sampling Period				Absolute Pressure, mm Hg								
Beginning		Ending		Minimum			Maximum			Mean	Standard	Number of
Date	Time	Date	Time	Value	Date	Time	Value	Date	Time	Value	Deviation	Values
222	1433	226	1217	714.40	225	1857	718.33	224	0245	716.51	0.90339	746
226	1242	228	1358	712.91	226	1831	718.49	228	1048	715.50	0.99725	548
228	1423	230	1157	714.85	228	0253	721.59	229	2304	718.86	1.8353	514
230	1222	232	1209	717.67	231	1803	720.93	230	1227	719.07	0.53936	551
232	1234	234	1046	717.50	234	0340	719.10	233	1209	718.25	0.63914E-32	523
234	1111	236	0954	716.84	234	1846	720.15	236	0949	718.42	0.63919E-32	533

USE OF PRECISION AND ACCURACY ESTIMATES FOR VALIDATION OF DATA

by

David T. Mage

Environmental Monitoring Systems Laboratory
U. S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

USE OF PRECISION AND ACCURACY ESTIMATES FOR VALIDATION OF DATA

David T. Mage

INTRODUCTION

A basic need in the presentation of a data set is a description of the precision and accuracy associated with the measurements. The definition of valid data for a given study is then determined by the precision and accuracy claimed for the data set. For example, a statement may be made that the data have a precision of 10%. Assuming that the errors are independent and normally distributed, one expects that 68% of the measurements are within ±10% and 95% of the measurements are within ±20% of the true value which is unknown. By this definition the 5% of the data in error by more than ±20% are not invalid since one expects from probability alone that runs of positive or runs of negative errors can occur. It is not productive and probably impossible to examine each datum point and determine an individual error associated with it. The approach being taken for the Community Health Air Monitoring Program (CHAMP) data base is to determine the precision and accuracy of the entire data set and not invalidate data except for known cause such as instrument failure. The alternative of eliminating data suspected of higher uncertainty in order to improve the precision of the remaining data set is counter productive in the context of a health study.

In a health study, the aerometric data are paired with health data. When aerometric data are invalidated, the associated health statistics are also removed from the analysis. Because the occurrence of the health indicator, such as an asthmatic attack, is relatively infrequent, the loss of the information significantly reduces the validity of the overall study. For this reason an approach which provides a large aerometric data base of moderate precision and accuracy is preferable to an approach which provides a reduced data base of higher precision and accuracy. The following sections describe the system of data validation currently being used by CHAMP.

ERROR ANALYSIS

When an aerometric analyzer is continually monitoring pollutant, several sources of potential error can influence the measurement. The ten sources of error given in Table 1 are discussed below.

1. Span Gas Analysis--This error covers the uncertainty in the process of preparing a known concentration of pollutant to provide an upscale reading.

This process may contain errors associated with preparation of a primary standard and subsequent analysis of a secondary or transfer standard to be used in the field. When dilution is necessary to attain the desired concentration, the errors in the flow measurements of standard and diluent air also contribute to the overall error. It is the belief of the author that this uncertainty is on the order of 5% ($\sigma_1 = .05$).

2. Zero Gas Impurity--The gas used to zero the instrument may contain some impurity. The presence of 0.1 ppm as opposed to 0.0 ppm represents potential error of 10% at the 1 ppm level and 1% error at the 10 ppm level. For the purpose of this analysis a low error of 2% ($\sigma_2 = .02$) is chosen since greatest concern is at or above the National Ambient Air Quality Standard (NAAQS) where this error is minimized.

3. Instrument Drift, Electronic--When the sample and reagent flows to the instrument are held constant, a constant input concentration produces a signal which fluctuates about a mean value. This "noise" in the output signal may be caused by electonic noise in the photomultiplier tube and other electrical components due to voltage, frequency and temperature fluctuations. The estimate of error for this effect is 3% ($\sigma_3 = .03$).

4. Instrument Drift, Flow Variations--After an instrument is calibrated, and with input concentration held constant, an increase or decrease of the sample flow rate will tend to cause the instrument to drift away from the equilibrium point. When reagent flows are also being mixed in a reaction chamber, such as ethylene flow in a chemiluminescent ozone analyzer, the fluctuations in reagent flows also influence the output signal. These flows may be influenced by fluctuations in atmospheric pressure and vacuum in the flow system. The overall effect of these variations of flow from the

TABLE 1

SOURCES OF ERROR IN AIR MONITORING DATA

		σ_i	σ_i^2
	MAJOR ERRORS		
1.	SPAN GAS ANALYSIS	.05	.0025
2.	ZERO GAS IMPURITY	.02	.0004
3.	INSTRUMENT DRIFT, ELECTRONIC	.03	.0009
4.	INSTRUMENT DRIFT, FLOW VARIATIONS	.06	.0036
5.	OPERATOR IMPRECISION	.04	.0016
			$\sum$ .0090
	MINOR ERRORS		
6.	NON-LINEARITIES OF SCALE	.02	.0004
7.	RESPONSE TIME	.02	.0004
8.	INTERFERENCES	.02	.0204
9.	PRESSURE-TEMPERATURE CORRECTIONS	.01	.0001
10.	DATA PROCESSING AND ROUND OFF	.01	.0001
	TOTAL EXPECTED ERROR $\sigma_T = \sqrt{\sum_{1}^{10} \sigma_i^2}$		$\sum$ .0014

σ_T = 10% FOR THE INDIVIDUAL DATUM POINTS calibration condition is taken to be 6% (σ_4 = .06).

5. Operator Imprecision--The station operator in performing calibrations must adjust potentiometers and rotometers and perhaps read the mean of a fluctuating signal. A different operator repeating these procedures will arrive at a slightly different result for each of them. The resulting uncertainty, due to the human element, is estimated at 4% (σ_5 = .04).

6. Non-Linearities of Scale--A linear relation is usually assumed between voltage output and pollutant input. Slight non-linearities of scale are usually masked by the uncertainties in the measurements themselves. Where the scale appears to be linear an error of 2% in the linearity is almost indistinguishable, consequently this error of 2% is treated as a possibility which cannot be ignored (σ_6 = .02).

7. Response Time--Due to the finite response time of the instruments, a rising signal will lag and tend to be underestimated and a falling signal will lag and tend to be overestimated. These errors are felt to provide an error on the order of 2% in the measurements (σ_7 = .02).

8. Interferences--Variations in the atmospheric composition from the composition of the gas used to calibrate the instrument can cause errors. Common gases, besides the common pollutants, which fluctuate in the atmosphere are CO_2 and H_2O. These fluctuations can cause variations in output signal on the order of 2% (σ_8 = .02).

9. Pressure Temperature Correction--When data are corrected to standard conditions (25°C and 760 mm Hg), uncertainties in measured pressure and temperature can cause a slight error on the order of 1% (σ_9 = .01).

10. Data Processing and Round Off--In analog-digital conversions and vice versa, an error is created. When data are outputted a round off error also occurs. These errors are quite small and probably less than 1% (σ_{10} = .01).

The net result of all of these effects, assuming that the variances are additive, is an overall uncertainty of ±10%. This is interpreted as follows. If the atmospheres at 10 stations were all 10 ppm of some arbitrary pollutant, one would expect the mean of all 10 station measurements to be 10 ppm, 7 of them would be between 9 and 11 ppm, 2 of them would be in the range 8-9 ppm and 11-12 ppm, and 1 would be over 12 ppm or less than 8 ppm.

CHAMP VALIDATION CRITERIA

The CHAMP aerometric system is unique in that it measures and records the secondary flow parameters within the instrument simultaneously with the measurement of the primary parameter (pollutant concentration).

The fluctuations in the secondary parameters produce uncertainties in the measurements as described previously. When the fluctuations exceed their expected range, or limit of normal operation, two possible causes must be investigated. The first is pure randomness which means that the data are valid. The second cause may be a component failure, such as a clogged capillary tube, and this produces a bias in the data. Besides signalling the need to repair the instrument, the analyst has a cause for invalidating that portion of the data set where the instrument was operating out of the normal range.

As an example of the usage of secondary parameter data for determining precision, accuracy, and validity, the effect of ethylene flow within the Bendix chemiluminescent ozone analyzer is chosen for discussion. Figure 1 is a typical plot of the ozone flow variations within a CHAMP station. On the day shown the ethylene flow (FETH) varied from 25 cc/min to 26.3 cc/min. Figure 2 is a histogram of the fluctuations of FETH about the previous calibration setting of FETH during the 11-day period August 15 - August 25, 1977. The histogram has a mean of -0.03 cc/min and a standard deviation of 0.29 cc/min. The mean close to zero confirms that the fluctuations are not producing an appreciable bias, as expected. The standard deviation can be related to a standard error by examining how the output of an ozone analyzer varies with FETH.

Figure 3 shows the instrumental bias (Δ ppm O_3) as a function of ΔFETH at various ozone levels. In this case the instruments were on the 0.5 ppm scale, and were calibrated with a sample flow of 1000 cc/min and a FETH of 25 cc/min (bias = 0 by definition). When these data are normalized by dividing bias by original ppm value, the percentage changes at all four concentrations follow a common curve, Figure 4. In this case a linear fit to these data is not justifiable. At 25 cc/min the slope of the curve is +1.6% per cc/min. The standard deviation of 0.3 cc/min, therefore, corresponds to a standard error in ozone of 0.5% as shown on Table 2. The biases of the measurements to changes in vacuum (VAC), sample flow of ozone (SFO_3), sample

FIGURE 1

Variation of Ethylene Flow

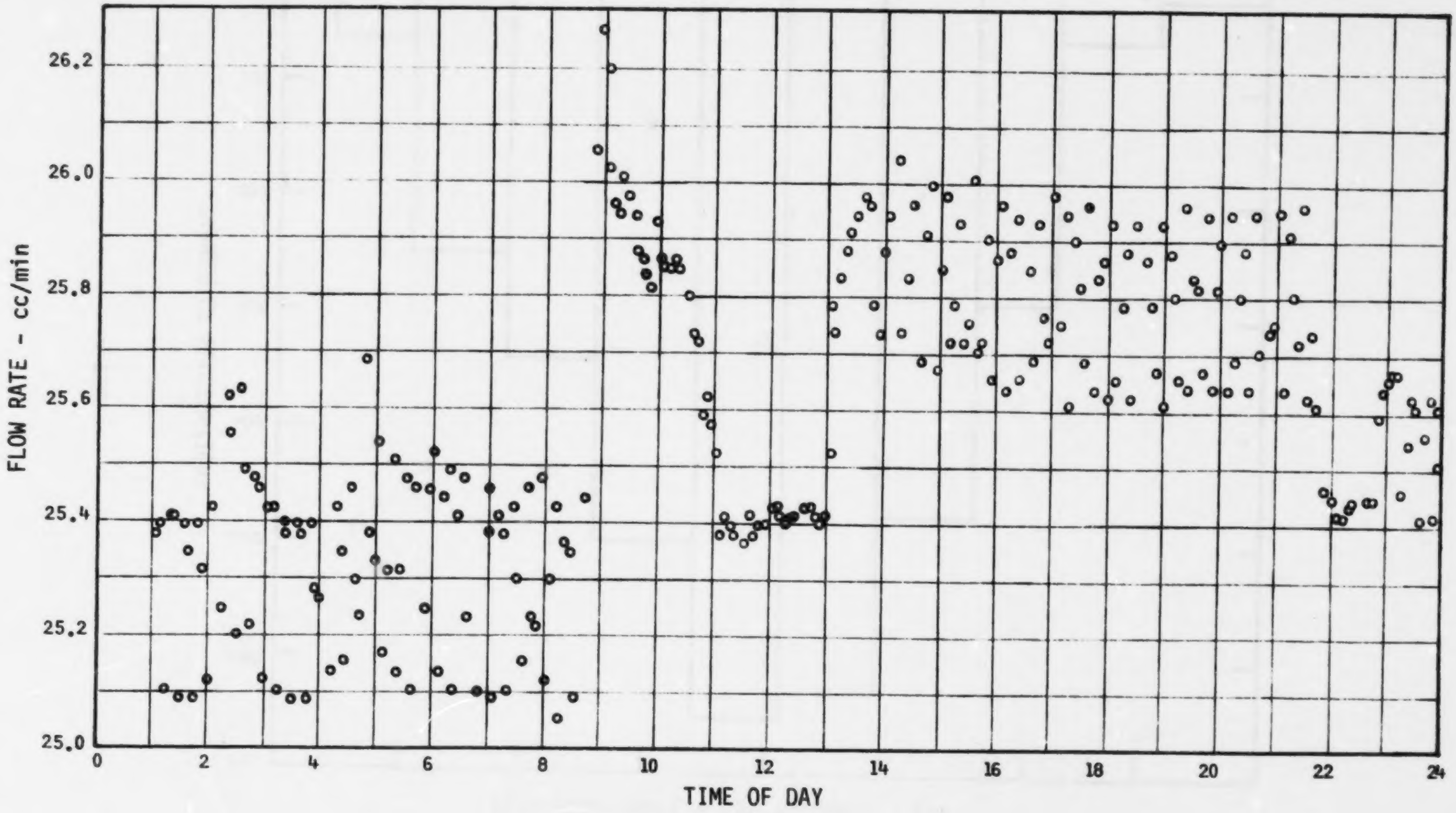
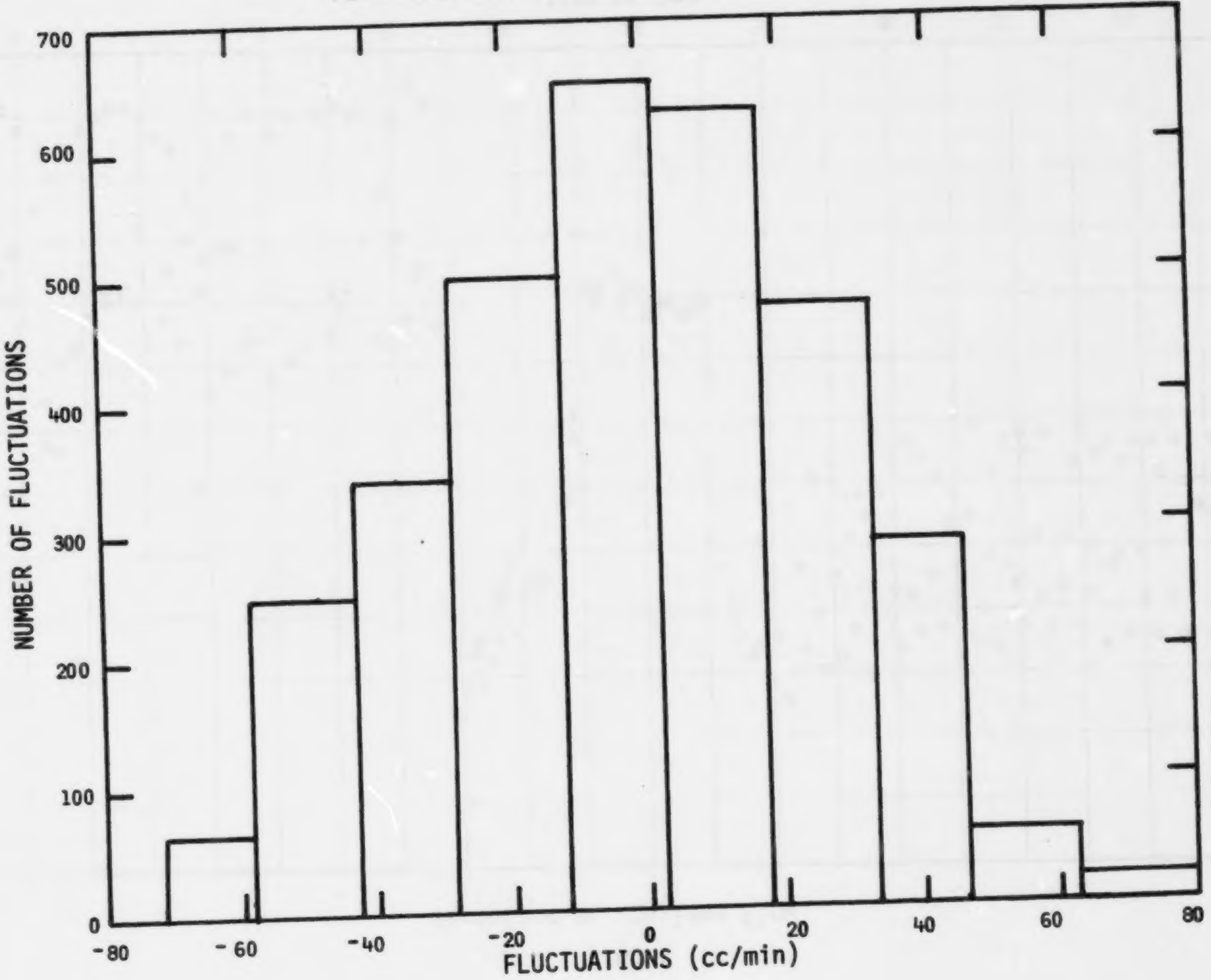


FIGURE 2

HISTOGRAM OF FLUCTUATIONS IN ETHYLENE FLOW

FETH FROM JDATE = 226 TO JDATE = 236



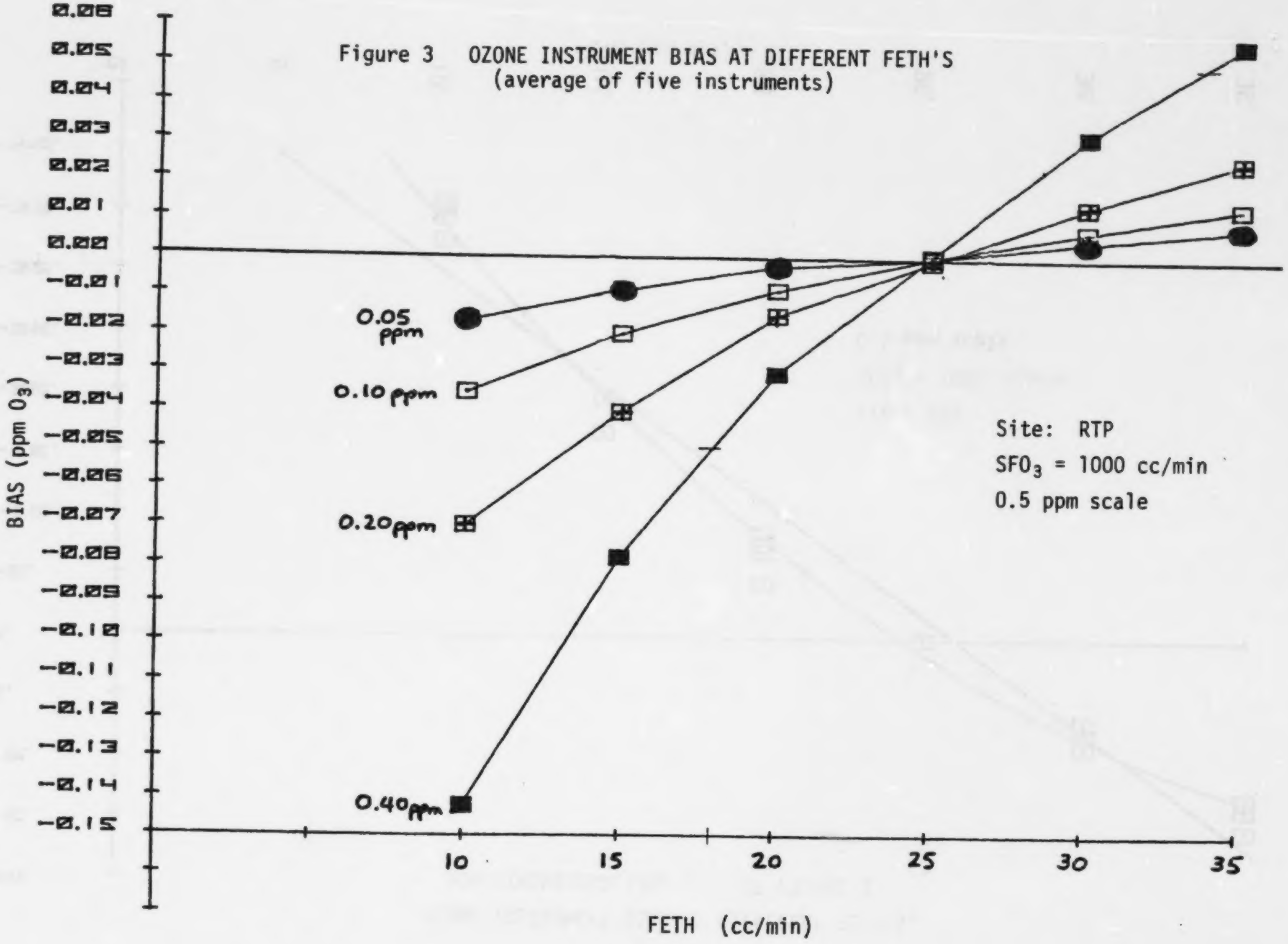


Figure 3 OZONE INSTRUMENT BIAS AT DIFFERENT FETH'S
(average of five instruments)

FIGURE 4

OZONE INSTRUMENT BIAS AT DIFFERENT FETH'S,
NON-DIMENSIONALIZED DATA OF FIGURE 3

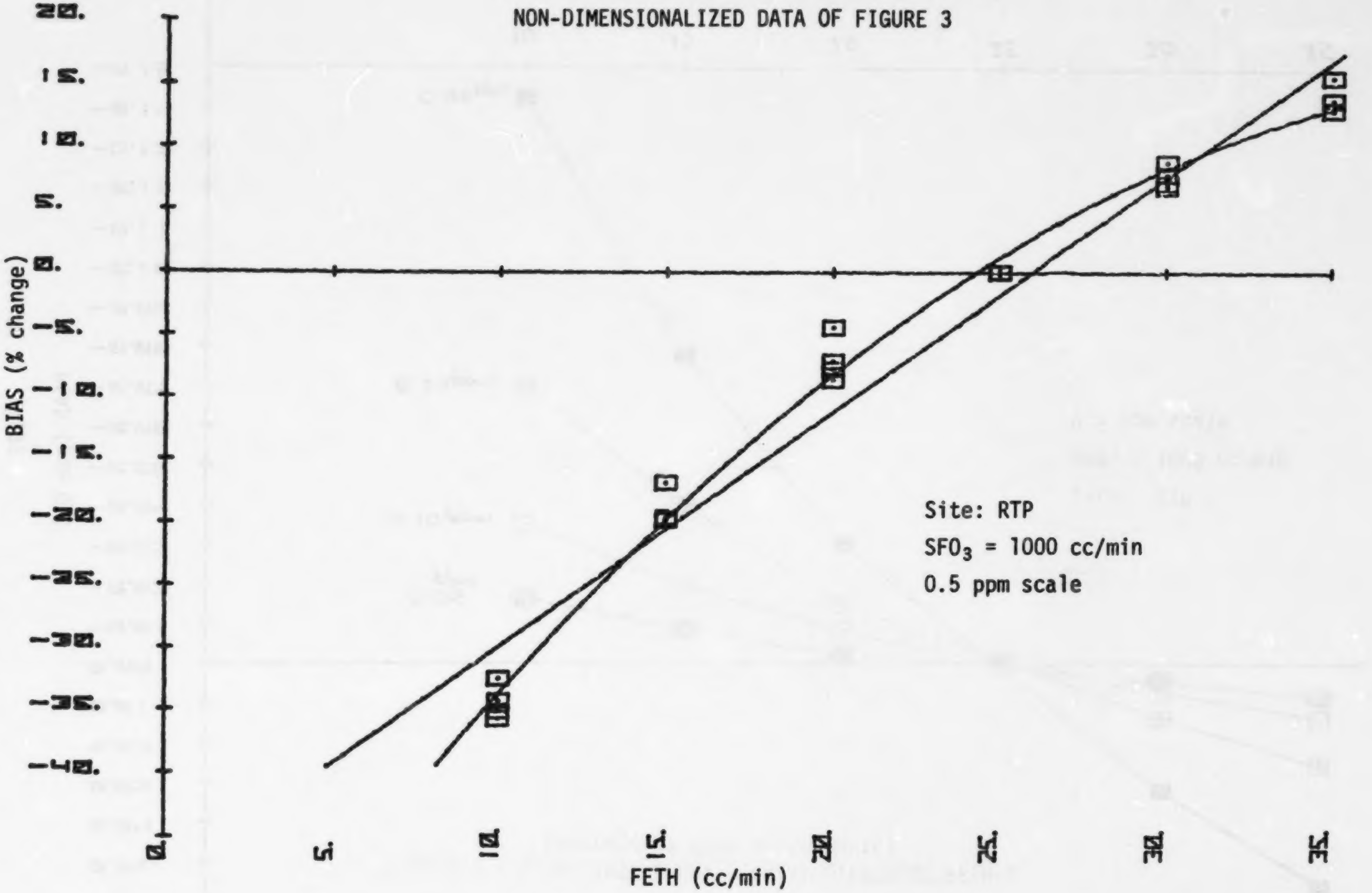


Table 2

Variations of Secondary Parameters

5-minute Averages

15-day Period 77222-77236

Upland, California

	SFNO	F02	VAC	FETH	SF03
Number of Observations	3294	3298	6461	3325	3187
Maximum	+4.1 cc/m	+1.0 cc/m	+13.3 torr	+.79 cc/m	+39.cc/m
Mean	- .08	+ .008	+ 0.88	_.03	-.07
Minimum	-5.7	-1.6	- 6.9	-.74	-95.
Standard Deviation	1.65 cc/m	0.48 cc/m	3.13 torr	0.29 cc/m	21.3 cc/m
Standard Error	1.6%	0.3%	2.5%	0.5%	0.5%

flow of NO (SFNO), and flow of air through the ozone generator (FO_2) are shown on Figures 5-8. These biases were used to compute the other standard errors given in Table 2.

DISCUSSION

The assumption developed in the preceding section is that the measurement errors from instrument-to-instrument are independent and normally distributed. When the same standard is used for repeated calibration of an instrument to provide a time series of measurements, the error in the analysis provides a bias. One of the functions of an audit, using an independent set of standards and observers, is to disclose whether a significant bias exists. In September of 1977, an audit of 46 instruments located at seven CHAMP stations was performed by the contract operator of the stations. The audit procedure was not truly independent since the transfer standards in station use were originally compared to the primary standard used in the audit. Of the 46 instruments, seven of one type showed a consistent bias indicating a problem with the audit procedure. The remaining 39 instruments showed an expected positive and negative scatter about the audit values.

In order to test the hypothesis that these 39 results are normally distributed, the results are plotted on normal probability paper at frequencies corresponding to their rank, lowest to highest, divided by the total (39) plus one. Each datum point represents the average of four span results at approximately 20%, 40%, 50%, and 80% of full scale. These four values of deviation are not independent since the same operator used the same standards for each one. However, the set of 39 averages are mutually independent. The mean of the deviation, μ, is -1.55% and the standard deviation, σ, is 5.5%. The maximum difference between the frequencies predicted by the normal distribution, $N(\mu,\sigma)$, and the data points is 6%. This corresponds to a Kolmogorov-Smirnov statistic of 0.06 which indicates that the hypothesis of normality for the distribution cannot be rejected at the 5% level. If an independent auditor performed the audit with independent standards, a standard deviation larger than 5.5% would be expected, probably on the order of 10 to 15%. The results of independent CHAMP audits are being analyzed in the manner described above and the results of the analyses will be reported with

FIGURE 5

NO_x INSTRUMENT BIAS AT DIFFERENT FO_2'S
(Average of five instruments)

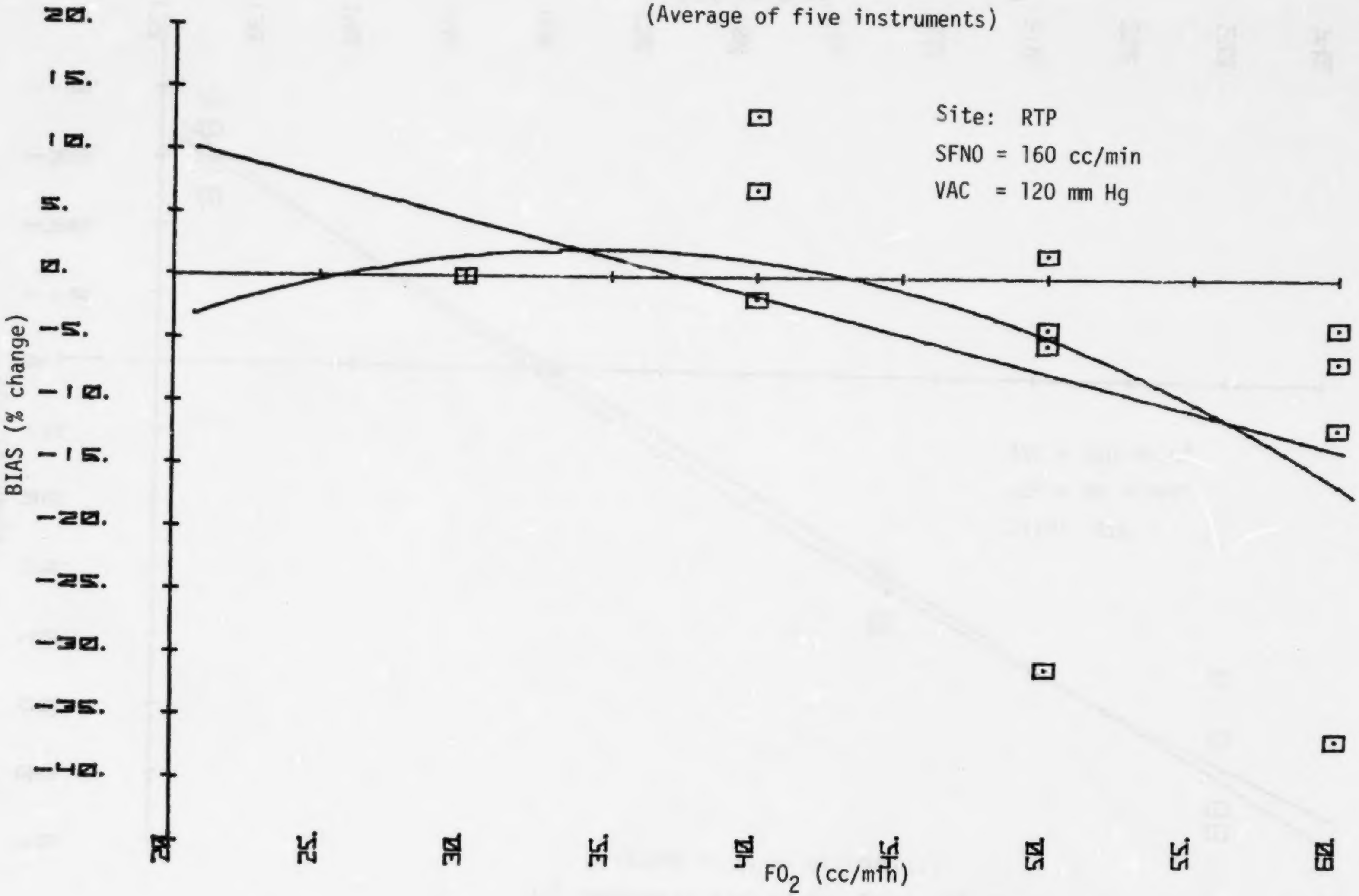


FIGURE 6

NO_x INSTRUMENT BIAS AT DIFFERENT SFNO'S

(Average of five instruments)

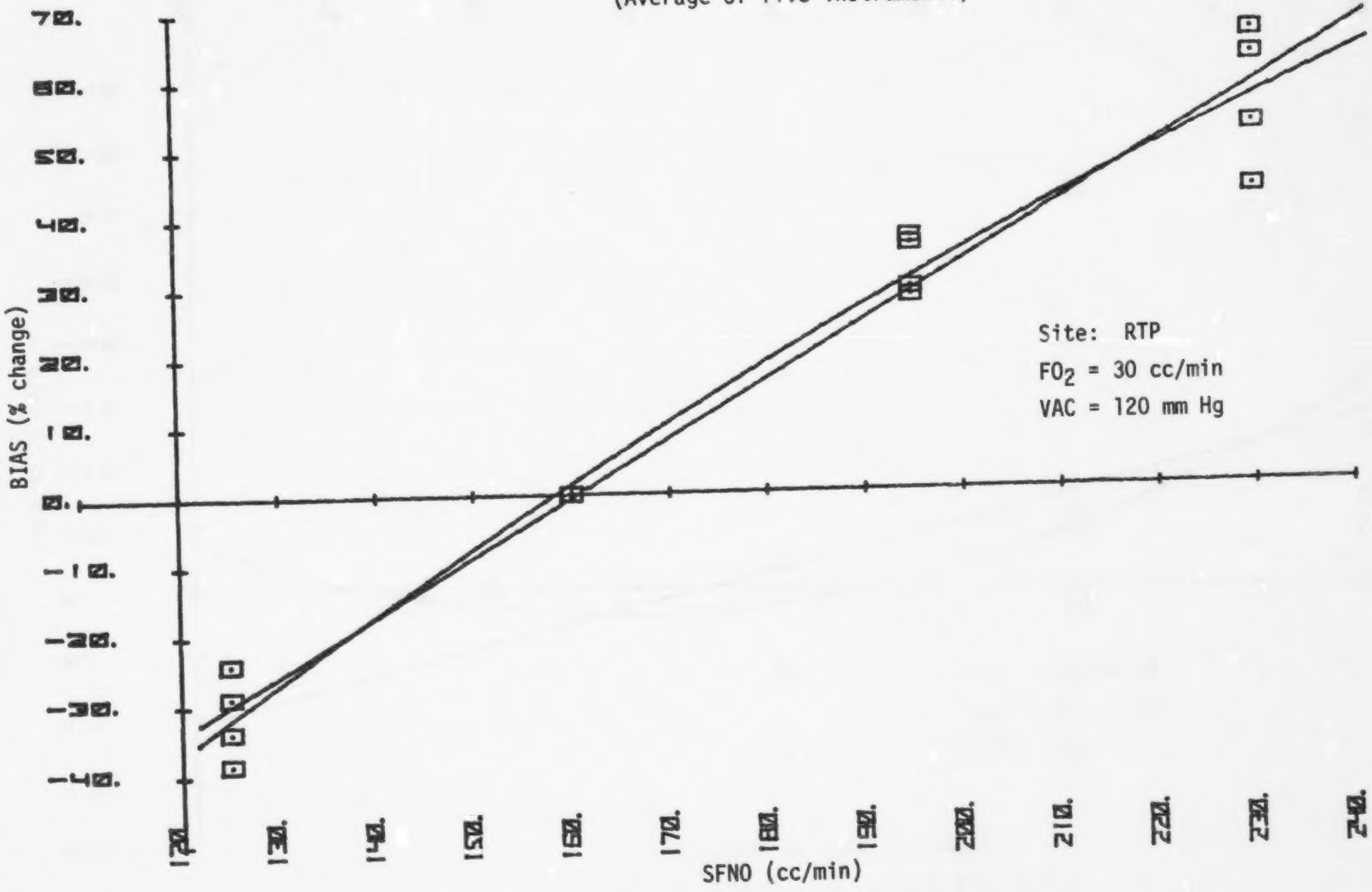
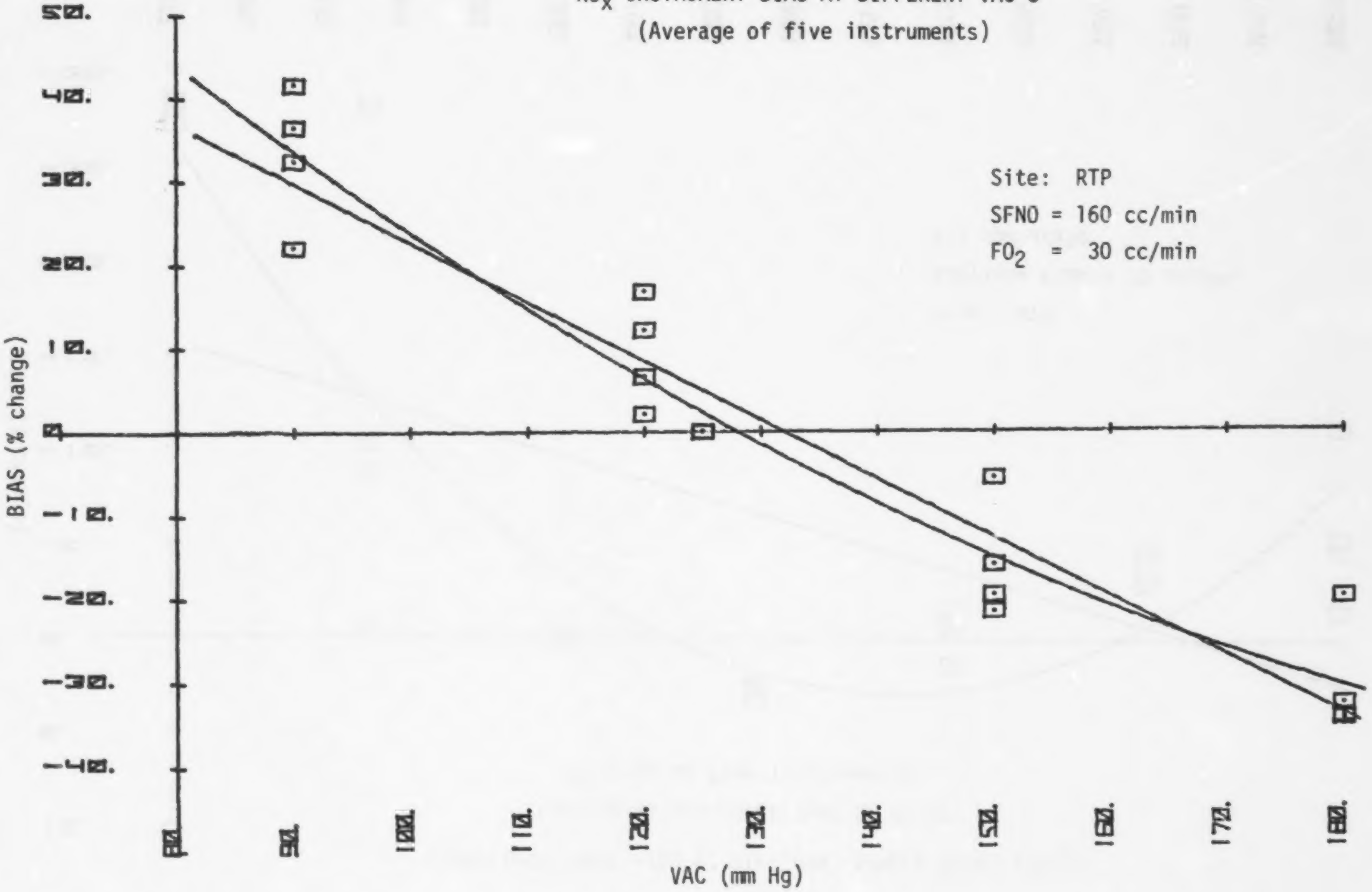


FIGURE 7

NO_x INSTRUMENT BIAS AT DIFFERENT VAC'S
(Average of five instruments)



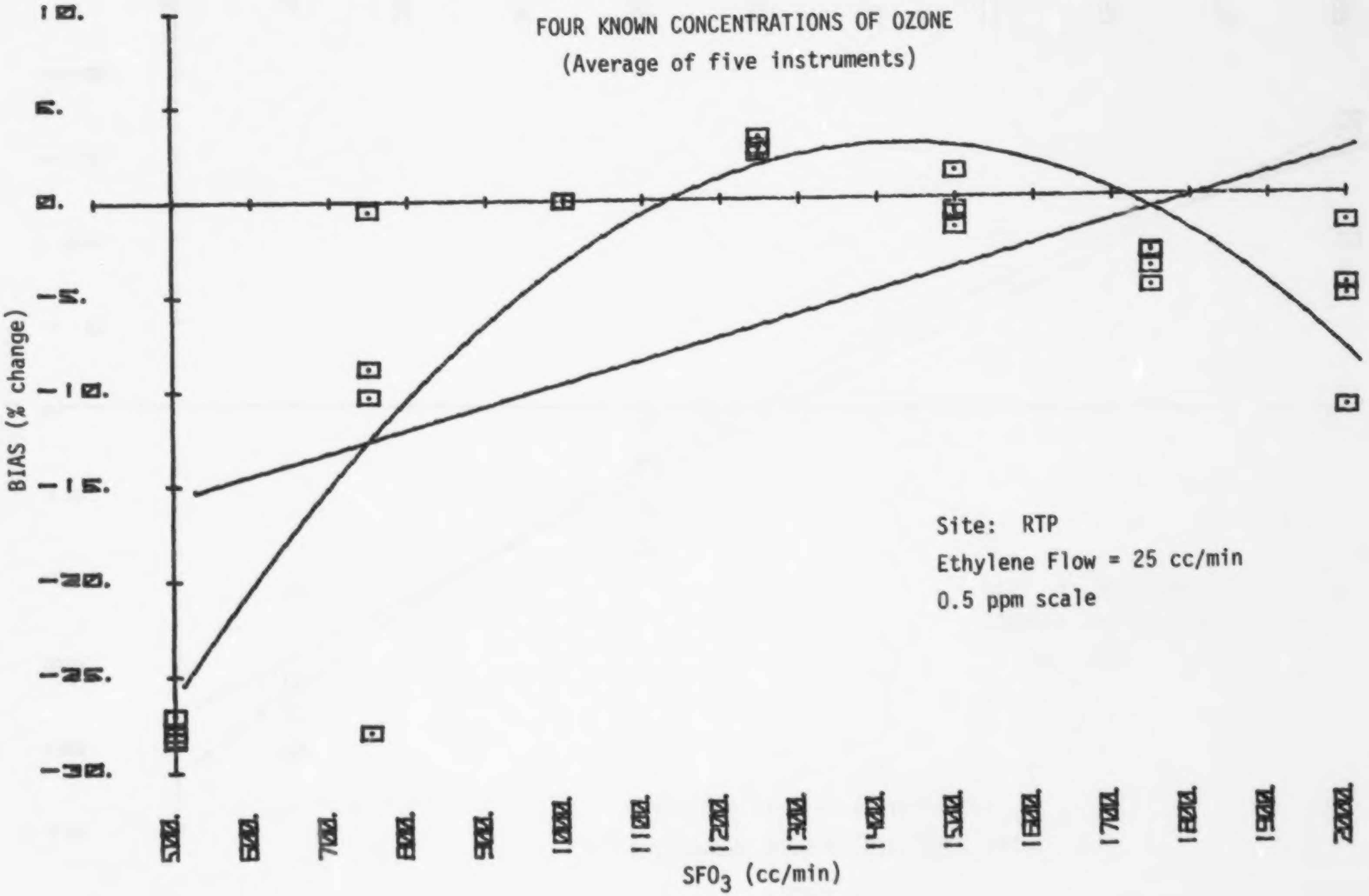
Site: RTP
SFNO = 160 cc/min
FO_2 = 30 cc/min

FIGURE 8

OZONE INSTRUMENT BIAS AT DIFFERENT SAMPLE FLOWS (SFO_3)

FOUR KNOWN CONCENTRATIONS OF OZONE

(Average of five instruments)



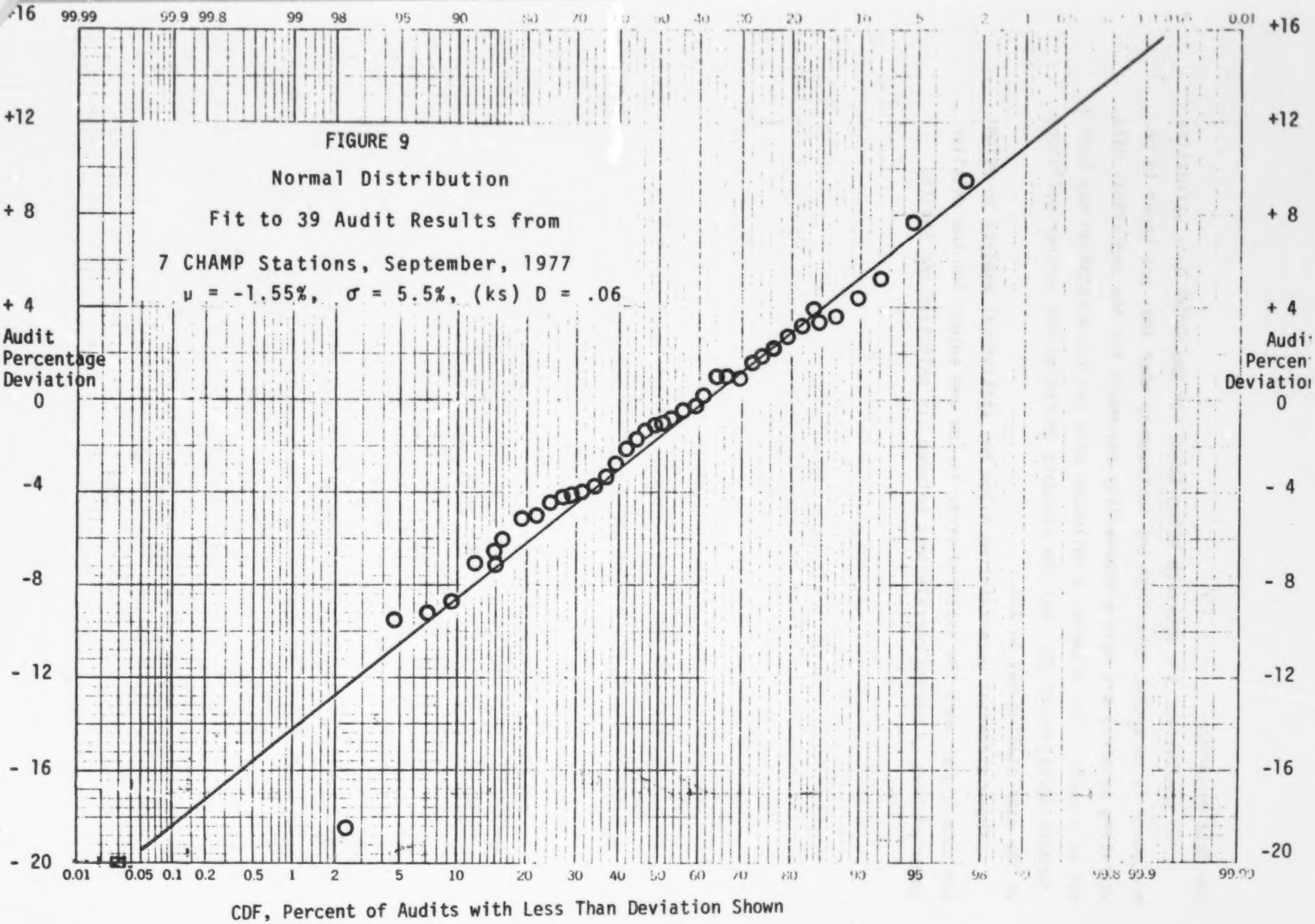


FIGURE 9

Normal Distribution

Fit to 39 Audit Results from

7 CHAMP Stations, September, 1977

$\mu = -1.55\%$, $\sigma = 5.5\%$, (ks) D = .06

the data elsewhere.

In conclusion, the approach being taken for the CHAMP data validation procedure is to accept data from the instruments when they are known to be operating properly and make a probability statement for the individual data set as a whole. For example, a pollutant data set for one station may have a standard deviation of 10% but the standard deviation for another pollutant at the same station may be 15%.

These different uncertainties allow the statistical analyst to weight the data higher when the expected error is low and adjust for the relative uncertainties in making correlations between air pollution and health.

VALIDATION SYSTEM USED IN THE ST. LOUIS REGIONAL AIR MONITORING STUDY (RAMS)

by

Robert B. Jurgens

Environmental Sciences Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

VALIDATION SYSTEM USED IN THE ST. LOUIS REGIONAL AIR MONITORING STUDY (RAMS)*

R.B. Jurgens+

Abstract

This paper describes the RAMS measurement system, screening categories of data validation, the RAMS automated validation system, the current status of special validation studies - visual validation and successive differences, and updates to the RAMS data base. The conclusion presents a generalized measurement system including quality control, data validation and feedback.

---

*Portions of this paper have been discussed in detail elsewhere (1) and will only be mentioned here for completeness.

+On assignment from National Oceanic & Atmospheric Administration. U.S. Department of Commerce, Rockville, Maryland 20852.

Introduction

The Regional Air Monitoring System (RAMS) is the ground-based aerometric network of the St. Louis Regional Air Pollution Study (RAPS). See references (2-3) for a discussion of the objectives, scope and accomplishments of RAPS.

The location of the 25 RAMS stations within the St. Louis metropolitan area is shown in Figure 1. The air quality, meteorological and solar radiation measurements within the RAMS network are listed in Table 1. Note that not all measurements are made at each station. Measurements began being recorded in mid summer 1974 and continued through June 1977. From April to June 1977 only stations 104, 106, 107, 111, 115, 121 and 125 were in operation. The approximate volume of data recorded during the network operation is 500 million values. Figures 2 and 3 show the data flow through the RAMS stations and through the central facility at Creve Coeur. Rockwell International Corporation was the prime contractor for the installation operation and maintenance of the RAMS network. A detailed description of RAMS can be found in references (4,5).

Quality Control & Data Screening

Data validity results from: 1) a quality control program designed to provide accurate data as it is measured and 2) a screening process to detect spurious values which exist despite the quality control process. The quality control program for the RAMS network is reviewed in (1) and (5). Detailed definition and discussion of the elements of quality control for air pollution measurement systems have been published in (6). The specific quality control activities relating to calibration, zero/span checks, status and analog checks associated with the gas analyzers are quite similar to those of the CHAMP program which are discussed in the preceding paper by Dr. Marvin Hertz.

Based on the experience of managing the data validation activities of RAMS we have developed a summary of screening techniques which is applicable for any continuous automated monitoring network (air pollution, water quality, etc.). These tests (Table 2) have been divided into three categories: 1) Operational, 2) Continuity and Relational and 3) A Posteriori. Discussion of screening tests within each category and their application in RAMS follows.

The first category, "Operational," contains checks which document the network instrument configuration and operating mode of the recording station. These checks, which in RAMS are part of the quality control program, include checks for station instrumentation, missing data, system analog and status sense bits, and instrument calibration mode. In addition to documenting system performance the checks are used to flag data in the RAMS archive. As designed, the RAMS data bank contains space for every potential measurement. For example, if an instrument is in calibration mode, the corresponding data slots will contain a "calibration" flag.

The second category, "Continuity and Relational," contains temporal and spatial continuity checks and relational checks between parameters which are based on physical and instrumental considerations or on statistical patterns of the data. A natural subdivision can be made between intrastation checks, checks which apply only to data from one station, and interstation checks, those which test the measured parameters for uniformity across the network.

Intrastation checks include tests for calibration drift (gas analyzers in RAMS), lower detectable limits, gross limits, aggregate frequency distributions, relationships, and temporal continuity.

The drift calculations, which are part of the quality control program, are discussed in the above references. Many measurement instruments have a threshold, or lower detection limit (LDL), below which their output is obscured by instrument noise. A standard practice adopted in RAMS is to replace values in this range (0.0 ± LDL) with +1/2 LDL. The LDL's for the gas analyzers and the wind speed sensor are the lower instrument limits listed in Table 3.

Gross limits, which in RAMS are used to screen impossible values, are based on the ranges of the recording instruments. These, together with the parametric relationships which check for internal consistency between values, are listed in Table 3. Setting limits for relationship tests requires a working knowledge of noise levels of the individual instruments. The relationships used are based on meteorology, atmospheric chemistry, or on the principle of chemical mass balance. For example, at a station for any given minute, TS cannot be less than $SO_2 + H_2S$ with allowances for noise limits of the instruments.

A refinement of the gross limit checks can be made using aggregate frequency distributions. With a knowledge of the underlying distribution, statistical limits can be found which have narrower bounds than the gross limits and which represent measurement levels that are rarely exceeded. A method for fitting a parametric probability model to the underlying distribution has been developed by Dr. Wayne Ott of EPA's Office of Research and Development (7). B.E. Suta and G.V. Lucha (8) have extended Dr. Ott's program to estimate parameters, perform goodness-of-fit tests, and calculate quality control limits for the normal distribution, 2- and 3-parameter lognormal distribution, the gamma distribution, and the Weibull distribution. These programs have been implemented on the OSI computer in Washington and tested on water quality data from STORET. This technique has not been implemented within RAMS.

Also, under intrastation checks are specific tests which examine the temporal continuity of the data as output from each sensor. It is useful to consider, in general, the types of atypical or erratic responses that can occur from sensors and data acquisition systems. Figure 4 illustrates graphically examples of such behavior, all of which have occurred to some extent within RAMS. Physical causes for these reactions include sudden discrete changes in component operating characteristics, component failure, noise, telecommunication errors and outages, and errors in software associated with the data acquisition system or data processing. For example, it was recognized early in the RAMS program that a constant voltage output from a sensor indicated mechanical or electrical failures in the sensor instrumentation. One of the first screens that was implemented was to check for 10 minutes of constant output from each sensor. Barometric pressure is not among the parameters tested since it can remain constant (to the number of digits recorded) for periods much longer than 10 minutes. The test was modified for other parameters which reach a low constant background level during night-time hours. SO_2 was generally at zero and no persistency check was applied against it.

A technique which can detect any sudden jump in the response of an instrument, whether it is from an individual outlier, step function or spike, is the comparison of successive differences of a measurement with predetermined control limits. These limits are determined for each parameter from the distribution of successive differences for that parameter. These differences will be approximately normally distributed with mean zero (and computed variance) when taken over a sufficiently long time series of measurements.

The type of "jump" can easily be identified. A single outlier will have a large successive difference followed by another about the same magnitude but of opposite sign. A step function will not have a return, and a spike will have a succession of large successive differences of one sign followed by those of opposite sign.

Though not implemented in Rockwell's data processing and validation program, TAPGEN, (partly because of an expected large increase in processing time on the PDP 11/40), or in EPA's data archiving programs (version 6.4) strong consideration has been given to this technique for potential applications in data screening and quality control checks. In 1976, the Data Management and Systems Analysis Section (DMSAS) awarded a contract to RTI to study validation procedures for the RAPS data bank (9) in which a major area of investigations was the use of minute successive differences. Use of successive differences in an ongoing special validation study will be discussed in a later section.

A number of interstation checks on meterological parameters are implemented in Rockwell's TAPGEN program. However, they have only been used for quality control of the RAMS system and not for validation (flagging) of RAMS data. These tests, which are shown in Table 4, are performed on hourly average data.

Another interstation check, the Dixon ratio test has been examined to determine its applicability for screening RAMS network outliers (1,9). Dr. Ty Hartwell, RTI, in an earlier session presented some results he has obtained using the Dixon ratio test on RAMS data. This test was never implemented into the RAMS data validation system.

Referring again to Table 2, the third screening category, "A Posteriori", was established to provide a mechanism for overriding the automated flagging schemes which have been implemented in the instrumentation at the remote sites and in the data screening module. From a review of station logs and preventive maintenance records, a knowledge of unusual events, or through visual inspection of data, it may be determined that previously valid data should be flagged as questionable. Conversely, it may be determined that previously invalid data should be validated by removing existing flags. An example of when data would be invalidated is when an instrument, such as a wind direction indicator, becomes misaligned or uncalibrated because of some non-linear or unknown reason. Removal of flags or revalidation can occur, for example, when the recording instruments function properly, but either the sense bit or analog status circuitry is known to have malfunctioned.

RAMS Automated Validation System

The screening tests used in validating RAMS data were largely developed and tested at RTP and then implemented in the St. Louis central facility computer for on-site, near-real-time processing. Through continued testing and modification the validation system evolved to its final version - version 6.4. All data archived by previous versions have been rearchived to this standard. Table 5 lists the causes and flags of screening tests while Figure 5 shows a flow diagram of the order in which the tests are applied.

Special Validation Studies

Special known problems have occurred on certain parameters from time to time. The origin of these problems can be traced to sensor failure, electrical transients, software bugs at RAMS stations and at the central facility, data acquisition hardware, etc. Despite the automated validation program these problems have lead to the archival of erroneous data. It should be noted that these problems have only effected a small percent of all data - estimated to be less than 1 percent of the total.

In an effort to locate, review and flag any remaining suspect data (known a priori or not) several studies have been initiated within DMSAS. Two major efforts involve a graphical review of hour average data and a computer study of minute successive differences:

Rams Hour Average Graphical Review. -- Table 6 lists the volume of data from the RAMS networks, the number of minute and hour plots and the number of microfiche (24x) required for plotting all RAMS data. The tremendously large number (70,000) of minute plots preclude a graphical review at this time interval. Therefore, a graphical review system using hour average data was developed. This system combines the use of computer graphics, interactive programs and computer files (lists) wherever possible to reduce the manual labor associated with the various tasks.

The steps in this study are shown in Figure 6. Computer-generated note books of hour average plots are reviewed by trained personnel for any suspect data. See Figure 7 for an example. The plots are also reviewed by a second individual. A consolidated list of dates and times is entered into the computer for input into an automatic retrieval of minute plots using the RAPS*GRAPHICS program. These plots from the RAMS minute archive are reviewed by DMSAS personnel. From the original review file a second computer disc file (preliminary update file) containing dates, time periods and suggested changes and flags is prepared. This list with corresponding minute/hour plots is forwarded to Rockwell for review and investigation of cause. With the concurrance of DMSAS and Rockwell the final output from the graphical review process is reached: an update file for the RAMS minute/hour archive.

Minute Successive Difference Study. -- Visual inspection of hour data will detect large discontinuties in time series plots of a measurement or uncorrelated traces between stations. However, if a few minutes of "spiky" data were recorded during an hour, the hour average may only be changed by a few percent. Since hour-to-hour variations in almost all RAMS parameters can normally be much larger than a few percent, small changes caused by errors in minute data will not be detected by visual observation.

To determine the quality of the minute archival data we have been applying a flagging procedure based on distribution functions of minute successive differences. This technique is based on the assumption that minute successive differences will approximate a normal distribution with mean zero. See Figure 8 for an example of a distribution function of ozone data from a five hour period.

The RTI study (9) has shown that for a given parameter sample standard deviations of minute successive differences are not constant over stations, time of day, or seasons. The functional form for Σ, which can be expressed as:

$$\Sigma \equiv \Sigma \text{ (parameter, date, time, station)}$$

is not known, however. Therefore, this study data flags have been chosen as $4^*\Sigma\text{max}$ where Σmax is the largest sample standard deviation found in the RTI study for a given parameter. These 4 sigma limits are listed below.

RAMS Variable	4 Sigma Limit
Windspeed (meters/sec.)	± 3.0
Temperature (°C)	± 0.7
Ozone (ppm)	± 0.010
CO (ppm)	± 1.97
Methane (ppm)	± 0.32
THC (ppm)	± 0.84
NO (ppm)	± .028
NO_X (ppm)	± .035
Total Sulfur (ppm)	± .022
SO_2 (ppm)	± .015

Flagged data (dates, times, station, etc.) are stored on the Univac 1110 and can be used as input in further analysis. Programs exist to automatically print and plot suspect data. An example of minute temperature data with a succession of outliers is shown in Figure 9. The corresponding hour average is circled in Figure 7.

Application of the minute succesive difference technique in a RAMS data base update module will permit recalculation of hour averages which contain significant amounts of erroneous spiky data.

RAMS Data Base Update

The visual validation and successive difference studies are part of a review of the RAMS data being conducted by DMSAS and Rockwell. The results of this review process will be an update file (dates, times, changes, flags, etc) and separate modules for an update program. Figure 10 shows this update process including review studies and specifically known problem areas. Underlying this review is the requirement that all changes will be documented and concurrance required as to probable cause of suspect data.

Monitoring System with Quality Control and Data Screening

Figure 11 illustrates the data validation process within the framework of a generalized monitoring system. Associated with sensor instruments and the data acquisition system are quality control blocks which contain those elements required for acquiring acceptable data: calibration, system status and sense bits, preventive maintenance, training and operation and maintenance documentation and records. Data processing and screening should take place soon after data acquisition to permit system feedback in the form of corrective maintenance, changes to control processes and even to changes in system design.

A control data set (or sets) should be created for use in software verification. When software changes are made, the control data set is processed and the output compared with previous versions. This is analagous to the recalibration of gas analyzers after maintenance. The control data set used in RAMS is 1-day's data from all sites.

The effectiveness of the data review process can be greatly enhanced by the use of graphics. Review of graphical displays of raw data permits a rapid continuity check of individual time series and a visual correlation of network data. Graphics naturally augments automated data validation techniques which are necessarily based on a priori knowledge of system performance characteristics, expected magnitude and variations in recording levels, etc.

A monitoring system is dynamic in nature - responding to changing hardware/software requirements and to variations in operating and maintenance procedures. On-site, near-real-time data review and allowance for feedback in system design can minimize the amount of lost or marginally acceptable data.

References

(1) Jurgens, R.B. and R.C. Rhodes,1976: Quality Assurance and Data Validation for the Regional Air Monitoring System of the St. Louis Regional Air Pollution Study. Proc. of the Conference on Environmental Modeling and Simulation, EPA 600/9-76-016, 730-735.

(2) Burton, C.S. and G. M. Hidy, 1974: Regional Air Pollution Study Program Objectives and Plans, EPA 630/3-75-009, 53 pp.

(3) Browning, R.H., 1977: (RAPS) Description and Status of the Data Measurements, Quality Assurance and Data Base Management System. (unpublished), 72 pp.

(4) Meyers, R.L. and J.A. Reagan, 1975: Regional Air Monitoring System at St. Louis, Missouri, International Conference on Environmental Sensing and Assessment, Paper 8-2, Lofc #75-37494, 4 pp.

(5) Hern, D.H. and M.H. Taterka, 1977: Regional Air Monitoring System Flow and Procedures Manual, EPA Contract DU 68-02-2093, 177 pp.

(6) Quality Assurance Handbook for Air Pollution Measurement Systems, 1976: Volume I, Principles, EPA 600/9-76-005, 365 pp.

(7) Ott, W.R., 1974: Selection of Probability Models for Determining Quality Control Data Screening Range Limits, Presented at 88th Meeting of the Association of Official Analytical Chemists, Washington, D.C., 6 pp.

(8) Suta, B.E. and G.V. Lucha, 1975: A Statistical Approach for Quality Assurance of STORET-Stored Parameters, SRI, EPA Control No. 68-01-2940, 8 pp.

(9) Hartwell, T. and F. Smith, 1977: Study of Two Data Validation Procedures for the RAPS Data Bank, RTI project 43U-1291-2, EPA Contract 68-02-2407, 46 pp.

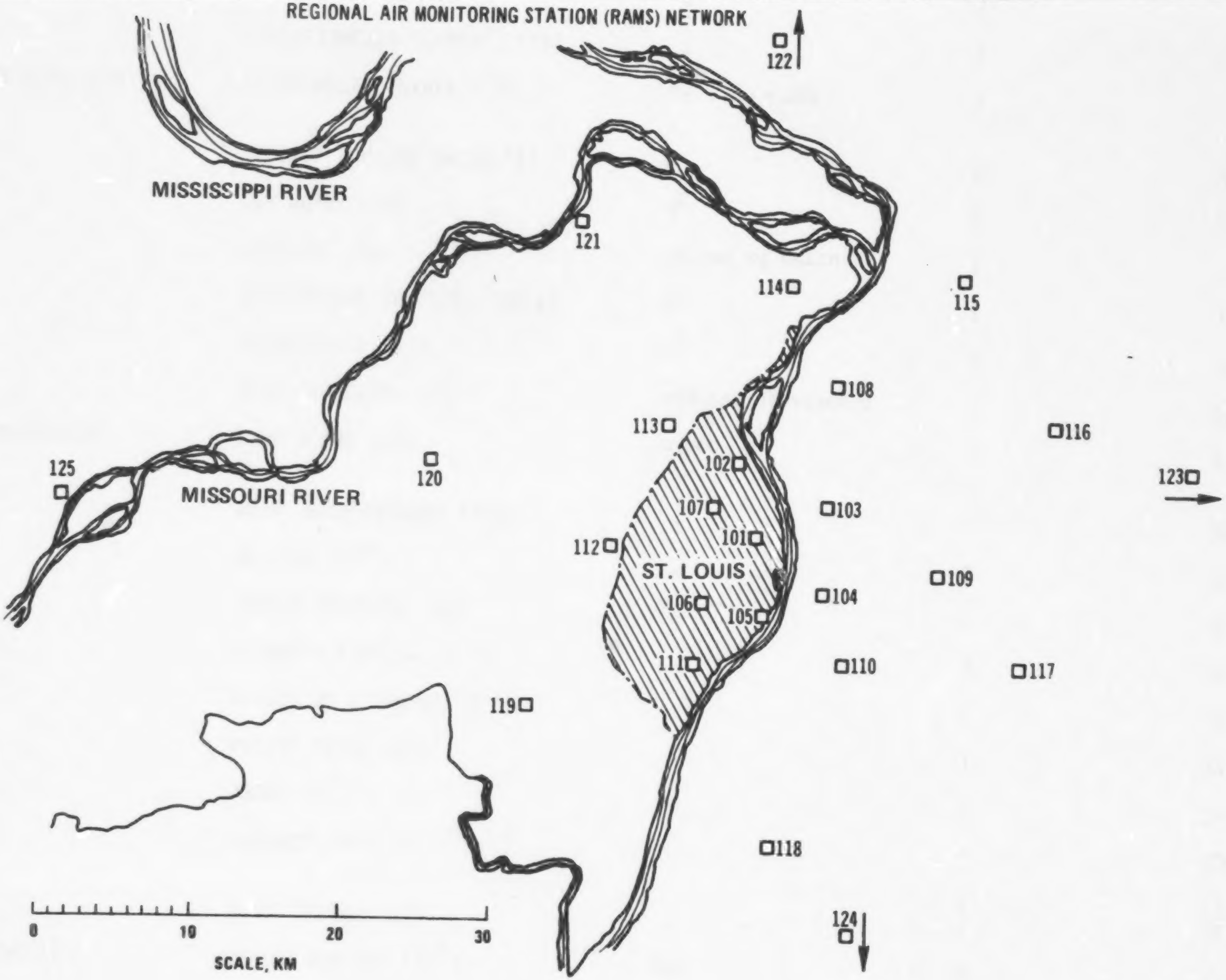

REGIONAL AIR MONITORING STATION (RAMS) NETWORK
MISSISSIPPI RIVER
MISSOURI RIVER
ST. LOUIS
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
0
10
20
30
SCALE, KM

Figure 1. Location of RAMS stations.

TABLE 1. RAMS NETWORK MEASUREMENTS

		UNITS	MEASUREMENT INTERVAL (min)	NUMBER STATIONS
AIR QUALITY:	SULFUR DIOXIDE (SO_2)	ppm	4	13
	TOTAL SULFUR (TS)	"	1 4	12 13
	HYDROGEN SULFIDE (H_2S)	"	4	13
	OZONE (O_3)	"	1	25
	NITRIC OXIDE (NO)	"	1	25
	OXIDES OF NITROGEN (NOX)	"	1	25
	NITROGEN DIOXIDE (NO_2)	"	1	25
	CARBON MONOXIDE (CO)	"	5	25
	METHANE (CH_4)	"	5	25
	TOTAL HYDROCARBONS (THC)	"	5	25
METEOROLOGICAL:	WIND SPEED (WS)	m/sec	1	25
	WIND DIRECTION (WD)	degrees of azimuth	1	25
	TEMPERATURE (TEMP)	^oC	1	25
	TEMPERATURE GRADIENT (DELT)	^oC	1	12
	PRESSURE (BP)	inches of mercury	1	7
	DEW POINT (DP)	^oC	1	25
	AEROSOL SCATTER (NEPH1,2)	m^{-1}	1	25
SOLAR RADIATION:	PYRANOMETER (PYRA1,2,3)	$CAL/cm^2 \cdot MIN$	1	6
	PYRHELIOMETER (PYRO1,...,9)	"	1	4
		"	1	4

RAMS DATA FLOW

A. RAMS STATION

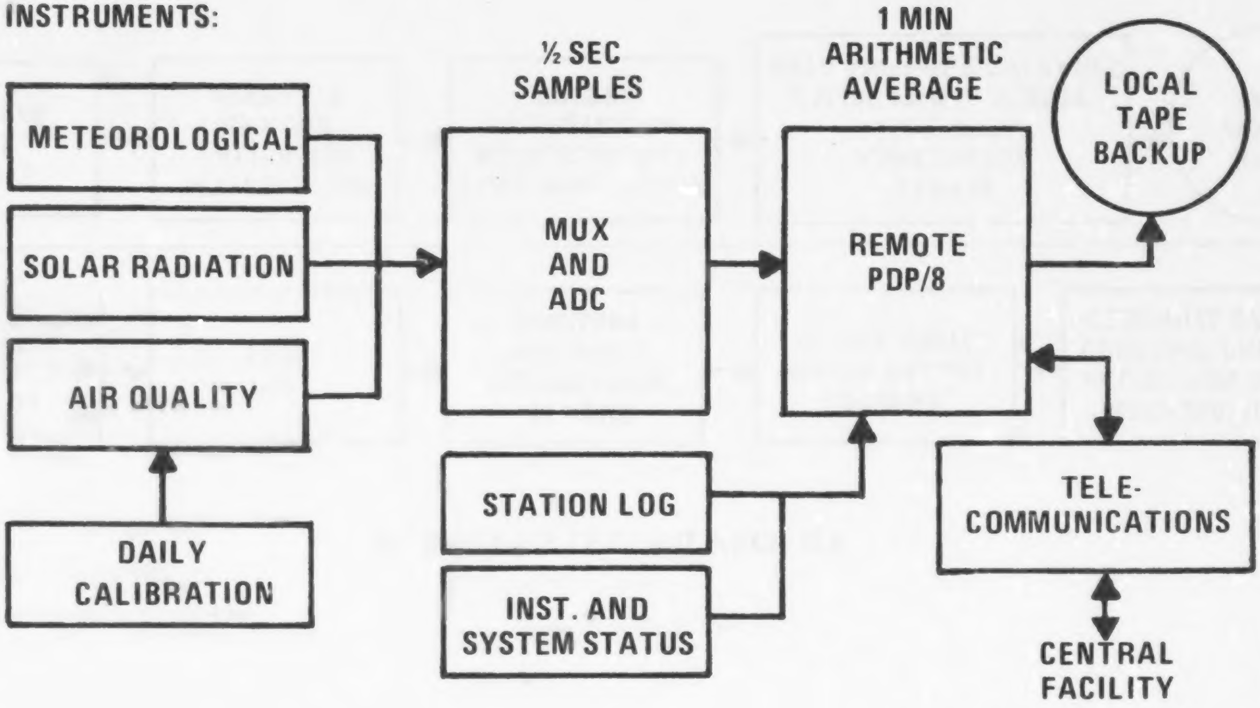
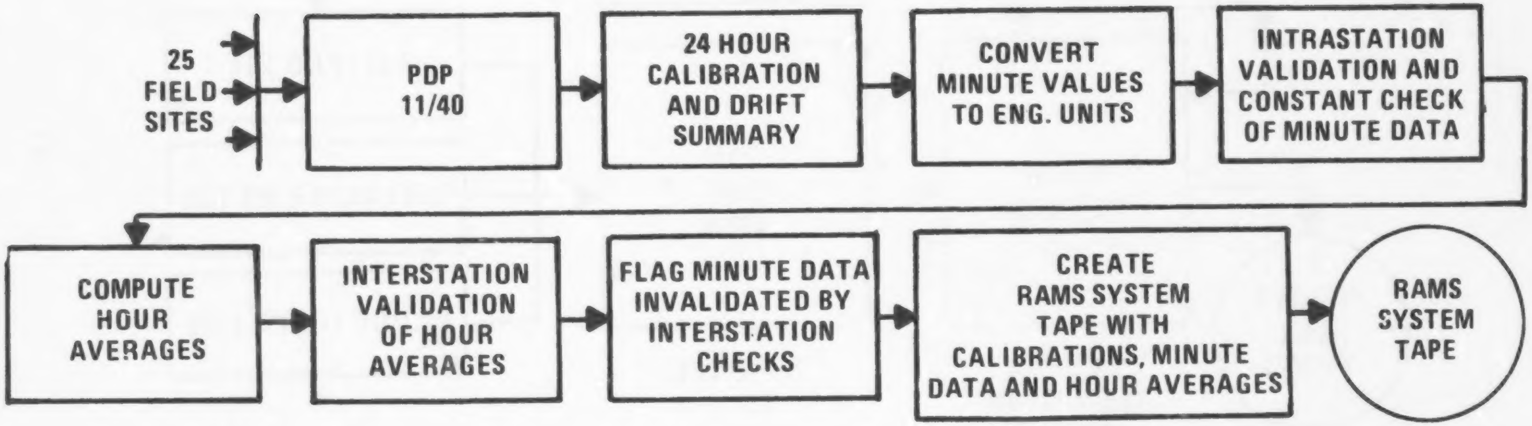

INSTRUMENTS:
METEOROLOGICAL
SOLAR RADIATION
AIR QUALITY
DAILY CALIBRATION
½ SEC SAMPLES
MUX AND ADC
STATION LOG
INST. AND SYSTEM STATUS
1 MIN ARITHMETIC AVERAGE
REMOTE PDP/8
LOCAL TAPE BACKUP
TELE-COMMUNICATIONS
CENTRAL FACILITY

Figure 2. RAMS data flow: RAMS station.



B. ST. LOUIS CENTRAL FACILITY

25 FIELD SITES
PDP 11/40
24 HOUR CALIBRATION AND DRIFT SUMMARY
CONVERT MINUTE VALUES TO ENG. UNITS
INTRASTATION VALIDATION AND CONSTANT CHECK OF MINUTE DATA
COMPUTE HOUR AVERAGES
INTERSTATION VALIDATION OF HOUR AVERAGES
FLAG MINUTE DATA INVALIDATED BY INTERSTATION CHECKS
CREATE RAMS SYSTEM TAPE WITH CALIBRATIONS, MINUTE DATA AND HOUR AVERAGES
RAMS SYSTEM TAPE

Figure 3. RAMS data flow: central facility.

TABLE 2. SCREENING CATEGORIES FOR AUTOMATED RECORDING NETWORKS

I. OPERATIONAL
  - NO INSTRUMENT
  - MISSING MEASUREMENT
  - STATUS
  - CALIBRATION

II. CONTINUITY AND RELATIONAL
  A. INTRA-STATION
    - CALIBRATION DRIFT
    - LOWER DETECTABLE LIMITS
    - GROSS LIMITS
    - AGGREGATE FREQUENCY DISTRIBUTIONS
    - RELATIONSHIP AMONG PARAMETERS
    - TEMPORAL CONTINUITY
      - CONSTANT OUTPUT
      - SUCCESSIVE DIFFERENCE

  B. INTER-STATION
    - METEOROLOGICAL NETWORK UNIFORMITY
    - STATISTICAL OUTLIERS
    - DIXON RATIO

III. A POSTERIORI
  - REVIEW OF STATION LOG
  - UNUSUAL EVENTS OR CONDITIONS
  - VISUAL INSPECTION OF DATA

TABLE 3. GROSS LIMITS AND RELATIONAL CHECKS

PARAMETER	INSTRUMENTAL LIMITS		INTERPARAMETER CONDITION
	LOWER	UPPER	
Ozone	.005 ppm	5 ppm	$NO*O_3 \le 0.01$
Nitric Oxide	.005 ppm	5 ppm	$NO - NO_X \le .002\ (NO)$
Oxides of Nitrogen	.005 ppm	5 ppm	$NO - NO_X \le .002\ (NO_X)$
Carbon Monoxide	.1 ppm	50 ppm	
Methane	.1 ppm	50 ppm	$CH_4 - THC \le .1\ (CH_4)$
Total Hydrocarbons	.1 ppm	50 ppm	$CH_4 - THC \le .1\ (THC)$
Sulfur Dioxide	.005 ppm	1 ppm	$SO_2 - TS \le .002\ (SO_2)$
Total Sulfur	.005 ppm	1 ppm	$SO_2 - TS \le .002\ (TS)$
Hydrogen Sulfide	.005 ppm	1 ppm	$H_2S - TS \le .002\ (H_2S)$
Aerosol Scatter	0.00001 m^{-1}	0.00099 m^{-1}	
Wind Speed	.27 m/s	22.2 m/s	
Wind Direction	0°	360°	
Temperature	-20°C	45°C	
Dew Point	-30°C	45°C	$DP - 1.0 \le T$
Temperature Gradient	-5°C	5°C	
Barometric Pressure	950 mb	1050 mb	
Pyranometers	-0.50	2.50 Langleys/min	
Pyrgeometers	0.30	0.75 Langleys/min	
Pyrehliometers	-0.50	2.50 Langleys/min	

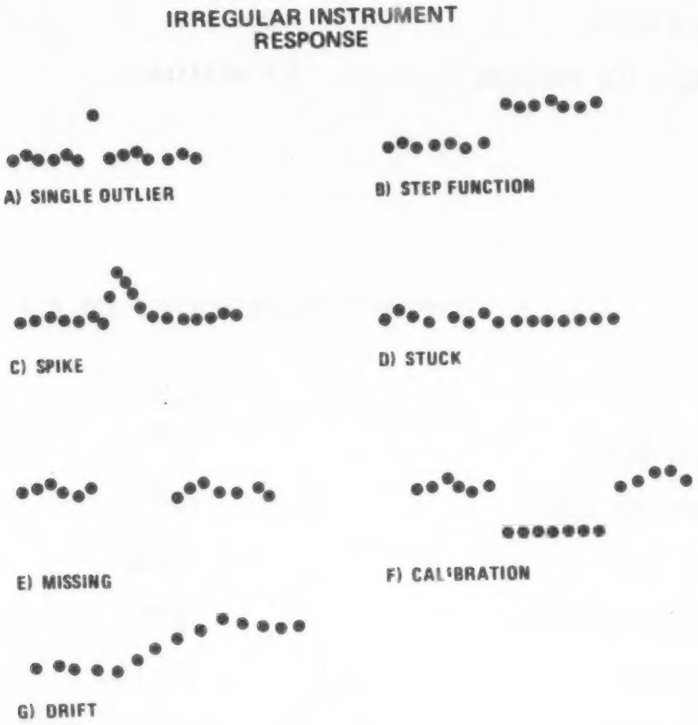

IRREGULAR INSTRUMENT RESPONSE
A) SINGLE OUTLIER
B) STEP FUNCTION
C) SPIKE
D) STUCK
E) MISSING
F) CALIBRATION
G) DRIFT

Figure 4. Irregular instrument response.

TABLE 4. MAXIMUM ALLOWABLE DEVIATIONS FROM NETWORK MEAN UNDER MODERATE WINDS (NETWORK MEAN > 4 m/sec)

WIND SPEED	2 m/sec OR mean/3 (WHICHEVER IS LARGER)
WIND DIRECTION	30°
TEMPERATURE	3°C
TEMPERATURE DIFFERENCE	0.5°C
DEW POINT	3°C
ADJUSTED PRESSURE	5.0 millibars

TABLE 5. RAMS DATA VALIDATION VERSION 6.4

	CAUSE	FLAG
1.	MISSING DATA	10^{37}
2.	CALIBRATION DATA	10^{35}
3.	EXCESS DRIFT	-VALUE
4.	FAILED RANGE TESTS	10^{34}
5.	LDL CHECKS	1/2 LDL
6.	STATUS ERROR	VALUE X 10^{-25}
7.	FAILED RELATIONAL TESTS	VALUE X 10^{32}
8.	FAILED TIME CONSTANT TESTS	VALUE X 10^{24}
9.	FAILED NETWORK TESTS	Q. A. REVIEW
10.	DATA MANAGEMENT OVERRIDE	VALUE X 10^{-15}

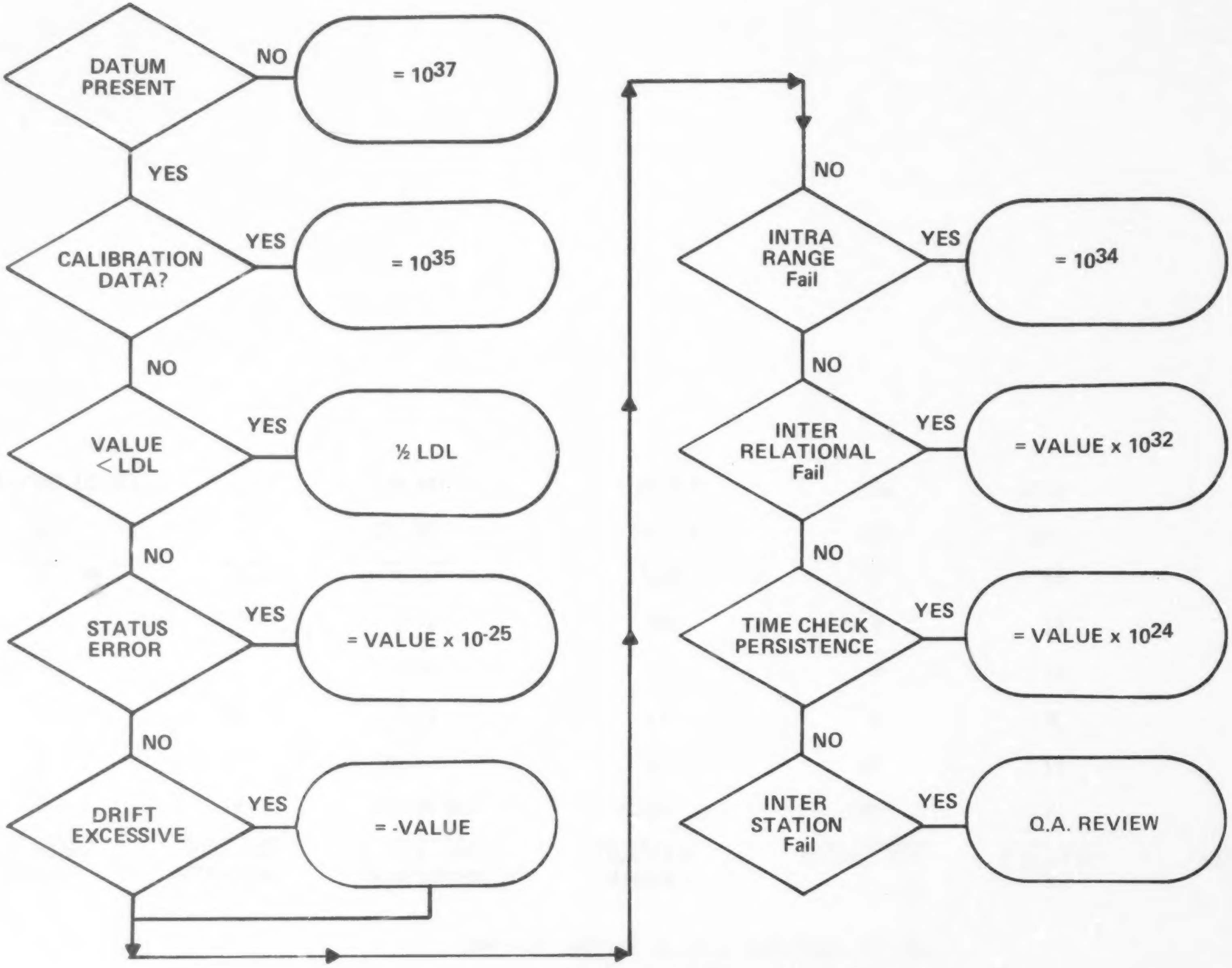

DATUM PRESENT
NO
= 10^37
YES
CALIBRATION DATA?
YES
= 10^35
NO
VALUE < LDL
YES
½ LDL
NO
STATUS ERROR
YES
= VALUE x 10^-25
NO
DRIFT EXCESSIVE
YES
= -VALUE
NO
INTRA RANGE Fail
YES
= 10^34
NO
INTER RELATIONAL Fail
YES
= VALUE x 10^32
NO
TIME CHECK PERSISTENCE
YES
= VALUE x 10^24
NO
INTER STATION Fail
YES
Q.A. REVIEW

Figure 5. Order of RAMS screening tests for minute data.

TABLE 6. VOLUME OF DATA FROM RAMS NETWORK

NUMBER PARAMETERS	RECORDING STATIONS	MEASUREMENTS PER YEAR	MINUTE PLOTS/YEAR	MICROFISCHE	HOUR PLOT/YEAR	MICROFISCHE
13	25	170.80 MIL.	23725	250	741	13
2	13	13.60	1900	20	76	2
1	12	6.31	730	8	38	1
1	7	3.68	365	4	19	1
1	6	3.15	365	4	19	1
2	4	4.20	730	8	38	2
TOTAL 20		202 MIL.	27.8 K	294	931	20
2 1/2 YEAR TOTALS:		500 MIL.	70.0 K	735	2330	20

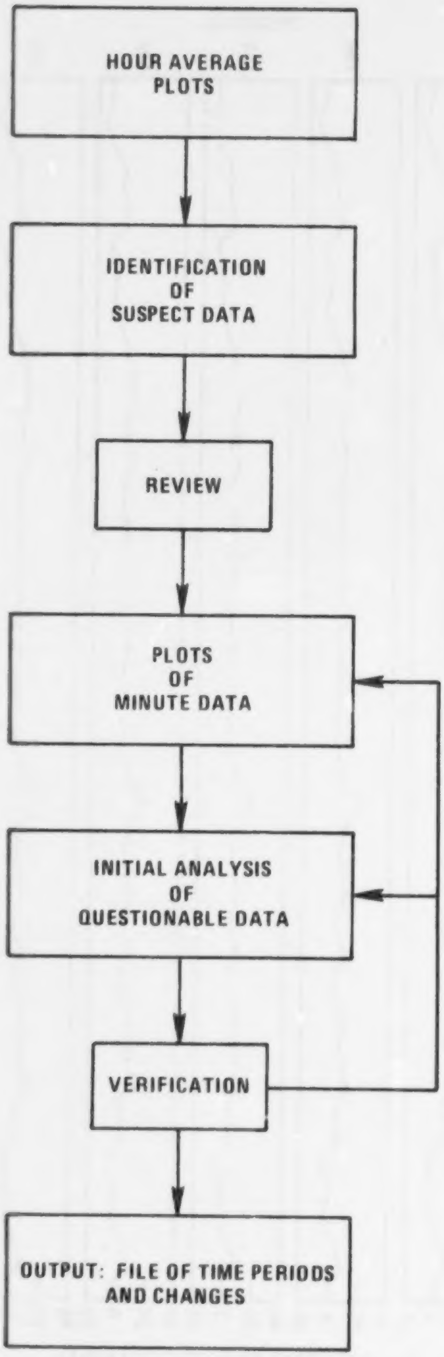

HOUR AVERAGE PLOTS
IDENTIFICATION OF SUSPECT DATA
REVIEW
PLOTS OF MINUTE DATA
INITIAL ANALYSIS OF QUESTIONABLE DATA
VERIFICATION
OUTPUT: FILE OF TIME PERIODS AND CHANGES

Figure 6. Visual review of RAMS hour average data.

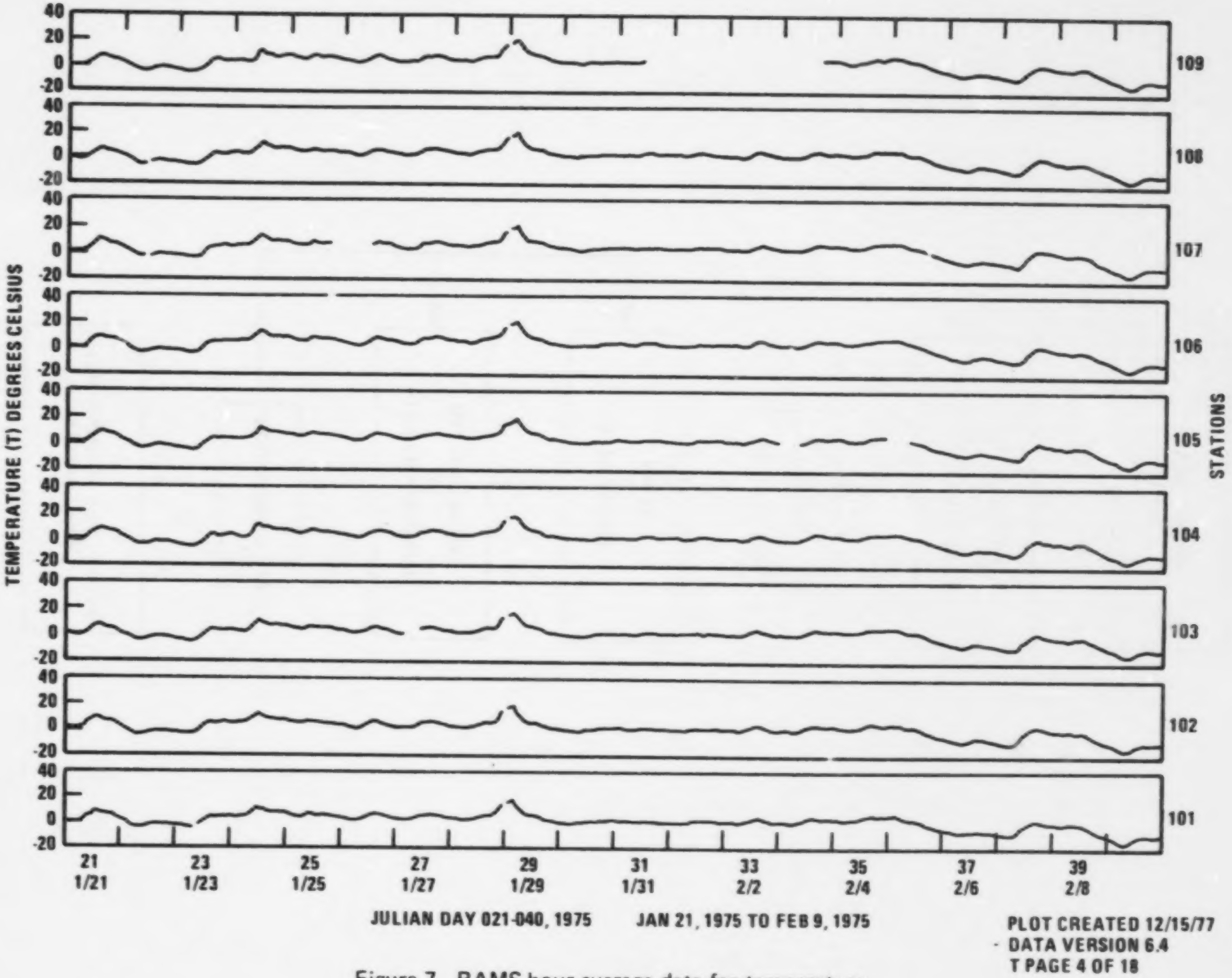


Figure 7. RAMS hour average data for temperature.

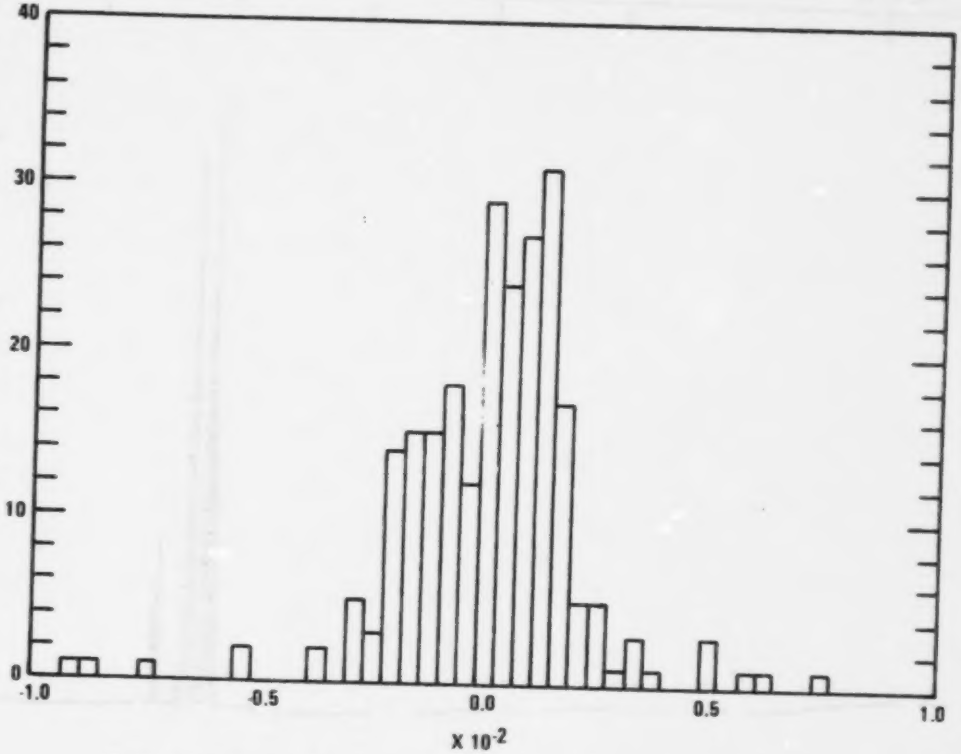


Figure 8. Ozone minute successive difference distribution function.

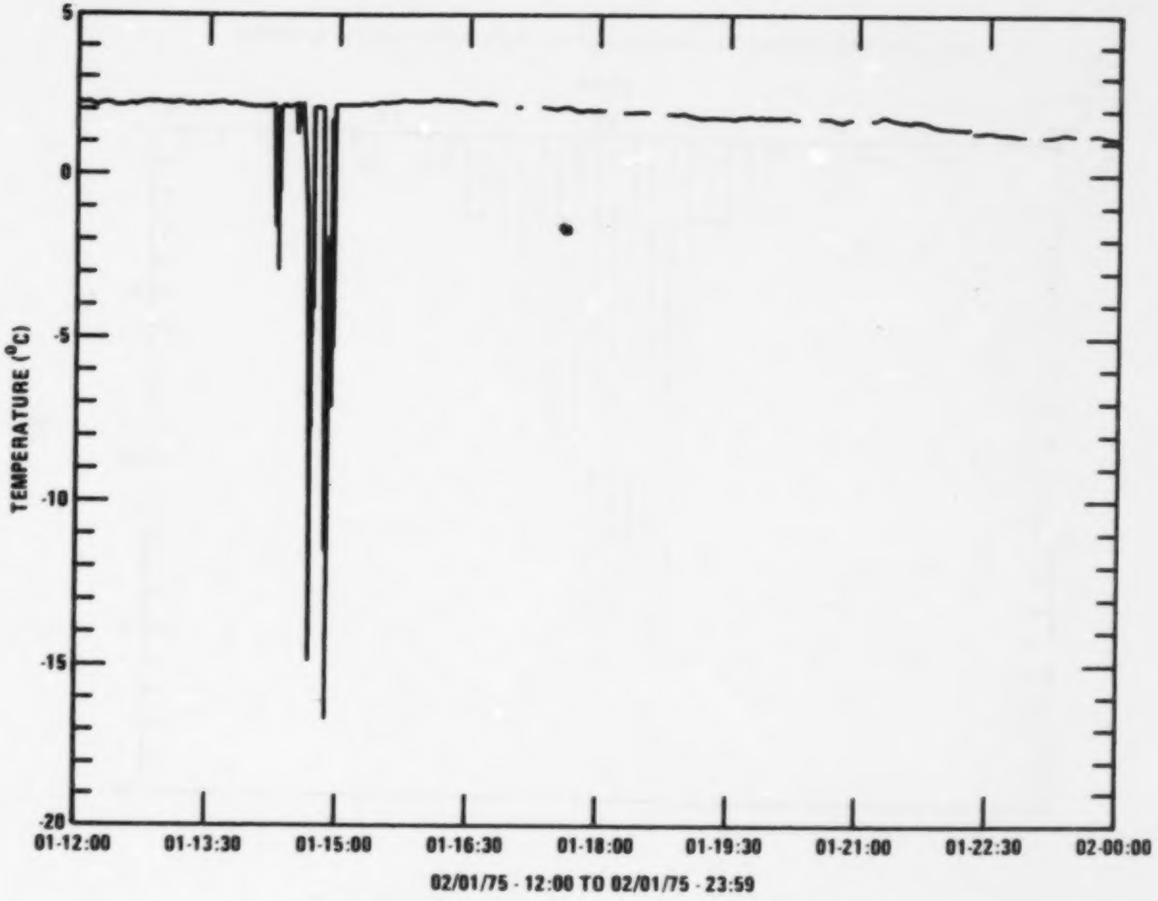


Figure 9. Minute temperature data from Station 103.

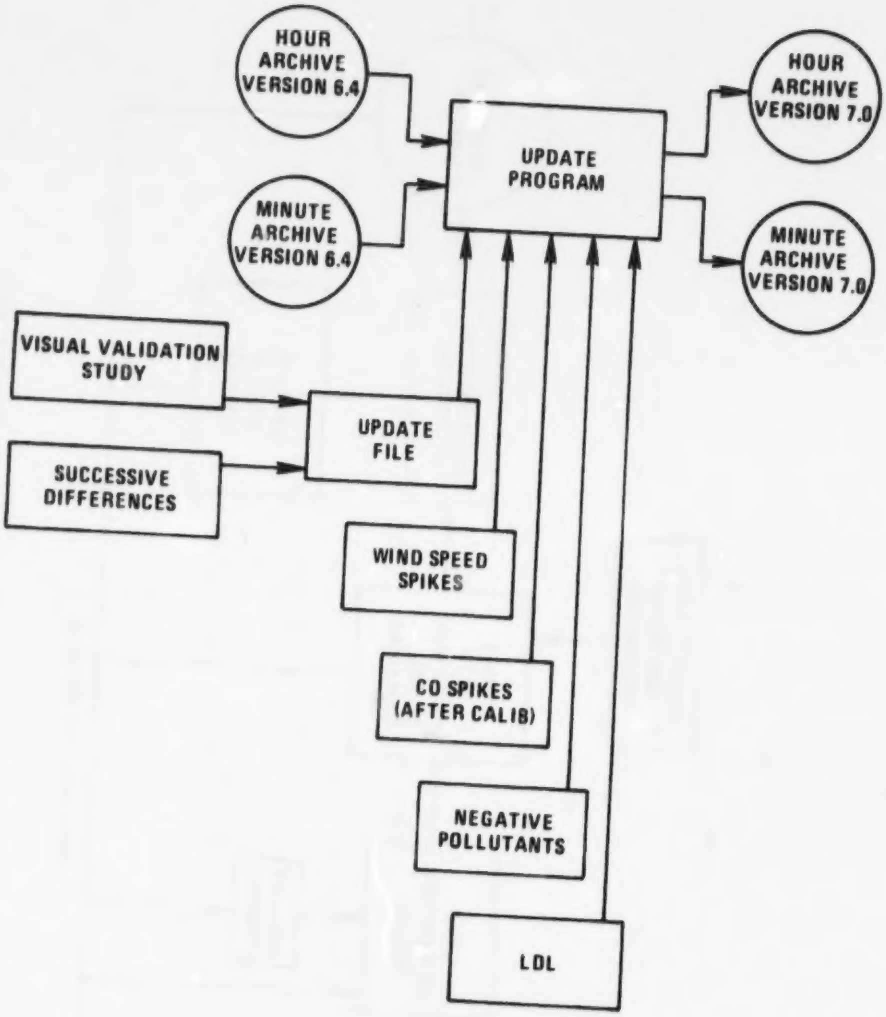

HOUR ARCHIVE VERSION 6.4
MINUTE ARCHIVE VERSION 6.4
UPDATE PROGRAM
HOUR ARCHIVE VERSION 7.0
MINUTE ARCHIVE VERSION 7.0
VISUAL VALIDATION STUDY
SUCCESSIVE DIFFERENCES
UPDATE FILE
WIND SPEED SPIKES
CO SPIKES (AFTER CALIB)
NEGATIVE POLLUTANTS
LDL

Figure 10. RAMS update.

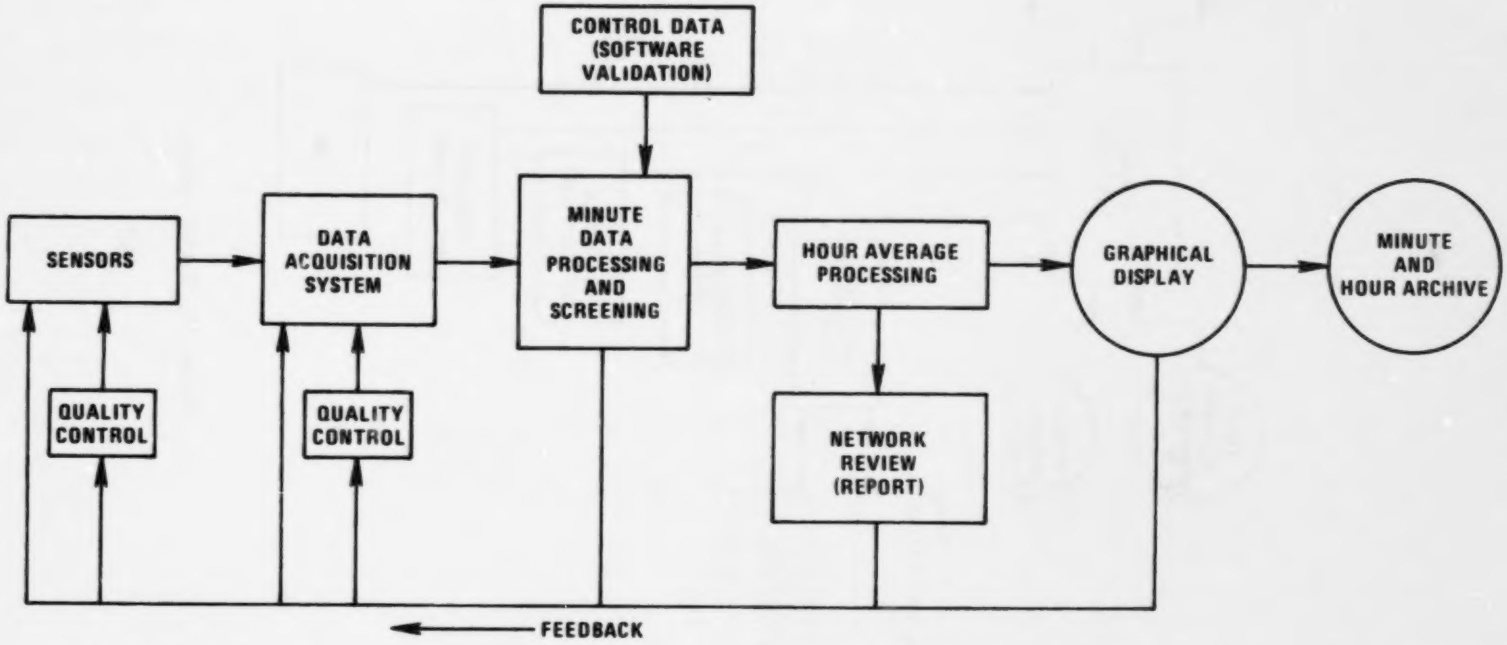

CONTROL DATA (SOFTWARE VALIDATION)
SENSORS
DATA ACQUISITION SYSTEM
MINUTE DATA PROCESSING AND SCREENING
HOUR AVERAGE PROCESSING
GRAPHICAL DISPLAY
MINUTE AND HOUR ARCHIVE
QUALITY CONTROL
QUALITY CONTROL
NETWORK REVIEW (REPORT)
FEEDBACK

Figure 11. Generalized monitoring system.

NAMES AND ADDRESSES:

PROGRAM

EPA/RTP INTERLABORATORY QUALITY ASSURANCE COORDINATING COMMITTEE

DATA VALIDATION CONFERENCE, SPEAKERS

DATA VALIDATION CONFERENCE, ATTENDEES

Conference on Data Validation
Research Triangle Park, North Carolina
November 4, 1977

The program that was distributed before the meeting is presented. Following this is an alternate schedule, which was used the day of the meeting.

CONFERENCE
ON
DATA VALIDATION

Environmental Research Center Auditorium
Highway 54 and Alexander Drive
Research Triangle Park, North Carolina

November 4, 1977

Sponsored by

ERC/RTP Interlaboratory Quality
Assurance Coordinating Committee

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Research and Development
Research Triangle Park, North Carolina

PROGRAM

8:00	**Registration**	
8:25	**Welcome**	Dr. J.K. Burchard *Senior ORD Official, RTP*
8:30	**Opening Remarks**	S. Hochheiser *EMSL*

GENERAL SESSION

D.J. von Lehmden
EMSL
Moderator

8:35	**What is Data Validation?**	R.C. Rhodes *EMSL*
8:45	**Validation Procedures Applied to In-Use Motor Vehicle Emission Data**	M.E. Williams *EPA, Ann Arbor*
9:15	**Computer Graphics in Data Validation**	Dr. R.H. Allen *COMP-AID*
9:45	**COFFEE BREAK**	
10:00	**Engineering Computations and Data Collection Formats Useful in Data Validation**	A.C. Nelson, Jr. *PEDCo*
10:30	**Regional Validation of State and Local Air Pollution Data**	T.H. Rose *EPA, Region IV*
11:00	**Use of Precision and Accuracy Estimates for Validation of Data**	Dr. D.T. Mage *HERL*
11:30 - 12:30	**LUNCH**	

AFTERNOON SESSIONS

CLASSROOM 2

	Special Applications	W.B. Kuykendal *IERL* *Moderator*
12:30	**Screening Checks Used by the National Climatic Center**	W.E. Klint *NCC, Asheville, NC*
12:55	**Data Validation for Upper Air Sounding Data and Emission Inventory Data**	J.H. Novak *ESRL*
1:20	**Validation of Continuous Stack Monitoring Data**	J.E. McCarley *OAQPS*
1:45	**Validation of Biomedical Data Through an On-Line Computer System**	L.D. Claxton *HERL*
2:10	**Data Validation Techniques Used in Mobile Source Testing**	C.D. Paulsell *EPA, Ann Arbor*

AUDITORIUM

	Statistical Techniques	G.G. Akland *EMSL* *Moderator*
12:30	**The Shewhart Control Chart Test for Screening 24-hour Air Pollution Measurements**	W.F. Hunt *OAQPS*
12:55	**Distribution Gap Test for Hourly Air Pollution Data**	Dr. T.C. Curran *OAQPS*
1:20	**Use of Statistical Sampling in Validating Health Effects Data**	C.P. Chamblee *HERL*
1:45	**Use of Successive Time Differences and Dixon Ratio Test for Data Validation**	Dr. T. Hartwell *RTI*
2:10	**Cluster Analysis as a Data Validation Technique**	Dr. H.L. Crutcher *Consultant*

2:35 - 2:50 **COFFEE BREAK**

(REMAINDER OF SESSIONS IN AUDITORIUM)

AUDITORIUM

Special Environmental Monitoring Studies		Dr. M.M. Bufalini *ESRL* *Moderator*
2:50	**Data Validation for the Los Angeles Catalyst Study (LACS)**	C.E. Rodes *EMSL*
3:25	**Validation Techniques Used in Continuous Air Monitoring Network (CHAMP)**	Dr. M.B. Hertz *HERL*
3:50	**Validation System Used in the St. Louis Regional Air Monitoring Study (RAMS)**	R.B. Jurgens *ESRL*
4:15	**Closing Comments**	S. Hochheiser *EMSL*

CONFERENCE ON DATA VALIDATION

Research Triangle Park, North Carolina

November 4, 1977

(Alternate Schedule)

AUDITORIUM

8:00	Registration	
8:25	Welcome	Dr. J.K. Burchard, Senior ORD Official, RTP
8:30	Opening Remarks	Seymour Hochheiser, EMSL
8:35	What is Data Validation?	R.C. Rhodes, EMSL

	CLASSROOM 2			AUDITORIUM	
General Session A		Dr. M.M. Bufalini ERSL, Moderator	General Session B		D.J. VonLehmden EMSL, Moderator
9:00	Engineering Computations and Data Formats Useful in Data Validation	A.C. Nelson, Jr. PEDCo	9:00	Validation Procedures Applied to In-Use Motor Vehicle Emission Data	M.E. Williams EPA, Ann Arbor
9:45	Coffee Break		9:45	Coffee Break	
10:00	Regional Validation of State and Local Air Pollution Data	T.H. Rose EPA, Region IV	10:00	Computer Graphics in Data Validation	Dr. R.H. Allen COMP-AID
10:45	Data Validation for the Los Angeles Catalyst Study (LACS)	C.E. Rodes EMSL	10:45	Use of Precision and Accuracy Estimates for Validation of Data	Dr. D.T. Mage HERL
11:30	LUNCH		11:30	LUNCH	
			AFTERNOON SESSIONS		
Special Applications		W.B. Kuykendal IERL, Moderator	Statistical Techniques		G.G. Akland EMSL. Moderator
12:30	Screening Checks Used by the National Climate Center	W.E. Klint NCC, Asheville, NC	12:30	The Shewhart Control Chart Test for Screening 24-hour Air Pollution Measurements	W.F. Hunt OAQPS

CLASSROOM 2

1:00	Data Validation for Upper Air Sounding Data & Emission Inventory Data	J.H. Novak ESRL
1:30	Validation of Continuous Stack Monitoring Data	J.E. McCarley OAQPS
2:00	Coffee Break	
2:15	Validation of Biomedical Data Through On-Line Computer System	L.D. Claxton, HERL
2:45	Data Validation Techniques Used in Mobile Source Testing	C.D. Paulsell EPA, Ann Arbor
3:15	Validation Techniques Used in Continuous Air Monitoring	Dr. M.B. Hertz HERL
4:00	Closing Comments	S. Hochheiser ESRL

AUDITORIUM

1:00	Distribution Gap Test for Hourly Air Pollution Data	Dr. T.C. Curran OAQPS
1:30	Use of Statistical Sampling in Validating Health Effects Data	C.P. Chamblee
2:00	Coffee Break	
2:15	Use of Successive Time Differences and Dixon Ratio Test for Data Validation	Dr. T. Hartwell RTI
2:45	Cluster Analysis as a Data Validation Technique	Dr. H.L. Crutcher Consultant
3:15	Validation System Used in the St. Louis Regional Air Monitoring Study (RAMS)	R.B. Jurgens ESRL
4:00	Closing Comments	R.C. Rhodes EMSL

Conference on Data Validation
Research Triangle Park, North Carolina
November 4, 1977

Members of the EPA/RTP* Interlaboratory Quality Assurance Coordinating Committee:

Mr. Seymour Hochheiser, Chairman
Assistant to the Director
EMSL
MD-75
RTP, NC 27711
Telephone: (919) 541-2106
FTS: 629-2106

Mr. Raymond C. Rhodes, Secretary
Quality Assurance Specialist
STAB/EMSL
MD-75
RTP. NC 27711
Telephone: (919) 541-2293
FTS: 629-2293

Mr. Ferris B. Benson
Quality Assurance Coordinator
HERL
MD-52
RTP, NC 27711
Telephone: (919) 541-2545
FTS: 629-2545

Dr. Marijon M. Bufalini
TPRO/ESRL
MD-59
RTP, NC 27711
Telephone: (919) 541-2949
FTS: 629-2949

Mr. William B. Kuykendal
Mechanical Engineer
IERL
MD-62
RTP, NC 27711
Telephone: (919) 541-2557
FTS: 629-2557

Mr. Darryl von Lehmden
Chemical Engineer
QAB/EMSL
MD-77
RTP, NC 27711
Telephone: (919) 541-2415
FTS: 629-2415

---

*Acronyms arranged alphabetically and used in this and the subsequent two sections.

EMSL - Environmental Monitoring and Support Laboratory
EPA - Environmental Protection Agency
ESRL - Environmental Sciences Research Laboratory
HERL - Health Effects Research Laboratory
IERL - Industrial Environmental Research Laboratory
MD - Management Division
NC - North Carolina
QAB - Quality Assurance Branch
RTP - Research Triangle Park
STAB - Statistical and Technical Analysis Branch
TPRO - Technical Planning and Review Office

Conference on Data Validation
Research Triangle Park, North Carolina
November 4, 1977

List of Speakers

Dr. Rod Allen
COMP-AID, Inc.
Box 12327
RTP, NC* 27709
Telephone: (919) 967-6376

Ms. Carolyn P. Chamblee
EPA/HERL
MD-55
RTP, NC 27711
Telephone: (919) 541-2348
FTS: 629-2348

Mr. Larry Claxton
EPA/HERL
MD-68
RTP, NC
Telephone: (919) 541-2518
FTS: 629-2518

Dr. Harold Crutcher
Consultant
35 Westall Ave.
Asheville, NC 28804
Telephone: (919) 253-2539
FTS: 672-0961

Dr. Thomas Curran
EPA/OAQPS
MD-14
RTP, NC 27711
Telephone: (919) 541-5351
FTS: 629-5351

Dr. Tyler Hartwell
RTI
Box 12194
RTP, NC 27709
Telephone: (919) 541-6453

Dr. Marvin Hertz
EPA/HERL
MD-56
RTP, NC 27711
Telephone: (919) 541-3124
FTS: 629-3124

Mr. William F. Hunt
EPA/OAQPS
MD-14
RTP, NC 27711
Telephone: (919) 541-5351
FTS: 629-5351

Mr. Robert B. Jurgens
EPA/ESRL
MD-80
RTP, NC 27711
Telephone: (919) 541-4545
FTS: 629-4545

Mr. William E. Klint
NOAA
Federal Building
Asheville, NC 28801
Telephone: (704) 258-2850, ext. 755
FTS: 672-0755

Dr. David T. Mage
EPA/HERL
MD-56
RTP, NC 27711
Telephone: (919) 541-3121
FTS: 629-3121

Mr. Joseph E. McCarley, Jr.
EPA/ESED
MD-13
RTP, NC 27711
Telephone: (919) 541-5245
FTS: 629-5245

Mr. A. Carl Nelson
PEDCo
Suite 201
5055 Duke Street
Durham, NC 27701
Telephone: (919) 688-6338

Ms. Joan Novak
EPA/ESRL
MD-80
RTP, NC 27711
Telephone: (919) 541-4545
FTS: 629-4545

Mr. C. Don Paulsell
EPA
2565 Plymouth Road
Ann Arbor, MI 48105
Telephone: (313) 668-4342
FTS: 374-8342

Mr. Charles E. Rodes
EPA/EMSL
MD-76
RTP, NC 27711
Telephone: (919) 541-3076
FTS: 629-3076

Mr. Thomas H. Rose
EPA/SB
College Station Road
Athens, GA 30605
Telephone: (404) 546-3489
FTS: 250-3489

Ms. Marcia Williams
EPA
2565 Plymouth Road
Ann Arbor, MI 48105
Telephone: (313) 688-4342
FTS: 374-8323

---

*See previous section for definition of acronyms. Acronyms not used previously are defined as follows:
NOAA - National Oceanic and Atmospheric Administration
OAQPS - Office of Air Quality Planning and Standards
RTI - Research Triangle Institute
GA - Georgia
MI - Michigan

Conference on Data Validation
Research Triangle Park, North Carolina
November 4, 1977

List of Attendees

Gerald G. Akland
EPA/EMSL/STAB
MD-75
RTP, NC 27711
Tel: (919) 541-2346
FTS: 629-2346

Rod Allen
COMP-AID
P.O. Box 12327
RTP, NC 27709
Tel: (919) 967-6376

Joseph S. Ali
EPA/HERL
MD-55
RTP, NC 27711
Tel: (919) 541-2240
FTS: 629-2240

J. Anderson
Rockwell International
5529 Chapel Hill Blvd.
Durham, NC 27707
Tel: (919) 942-2407

D. W. Armentrout
PEDco
1499 Chester Road
Cincinnati, OH 45246
Tel: (513) 782-4700

James D. Ashworth
U.S. Army Corps of Engineers
P.O. Box 2127
Huntington, WV 25721
Tel: (FTS) 924-5694

Andy Berlin
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

John Boston
EPA/SDMO
MD-55
RTP, NC 27711
Tel: (919) 541-2337

Frank Briden
EPA/IERL
MD-60
RTP, NC 27711
Tel: (919) 541-2557
FTS: 629-2557

T. G. Brna
EPA/IERL
MD-61
RTP, NC 27711
Tel: (919) 541-2915
FTS: 629-2915

Steve Bromberg
EPA/QAB
MD-77
RTP, NC 27711
Tel: (919) 541-2273
FTS: 629-2273

Robert Browning
EPA/ESRL
MD-80
RTP, NC 27711
Tel: (919) 541-4545
FTS: 629-4545

Sam Bryan
EPA
Chapel Hill, NC 27514
Tel: (919) 541-2872
FTS: 629-2872

Marijon M. Bufalini
EPA/ESRL
MD-59
RTP, NC 27711
Tel: (919) 541-2949
FTS: 629-2949

Bob Burton
EPA/HERL
MD-52
RTP, NC 27711
Tel: (919) 541-1394
FTS: 629-1394

D. Calafiore
EPA/HERL
MD-54
RTP, NC 27711
Tel: (919) 541-2674
FTS: 629-2674

Jon Carpenter
EPA
Ann Arbor, MI 48105
Tel: (FTS) 374-4293

Tom Caldwell
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Susan S. Casada
Northrop Services, Inc.
P.O. Box 12313
RTP, NC 27709
Tel: (919) 549-0611

Carolyn Chamblee
EPA/HERL
MD-55
RTP, NC 27711
Tel: (919) 541-2518
FTS: 629-2518

Ronald Chambler
NCHS - DPB
Box 12214
RTP, NC 27709
Tel: (919) 541-4422
FTS: 629-4422

Jonn Chavy
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Larry Claxton
EPA/HERL
MD-68
RTP, NC 27711
Tel: (919) 541-2518
FTS: 629-2518

John Clements
EPA/EMSL
MD-77
RTP, NC 27711
Tel: (919) 541-2196
FTS: 629-2196

Wayne Clements
TVA
345 Evans Bldg.
Knoxville, TN 37902
Tel: (615) 632-4579

William M. Cox
EPA/OAQPS
MD-14
RTP, NC 27711
Tel: (919) 541-5312
FTS: 629-5312

C. L. Cox, Jr.
EPA/ADM
MD-30
RTP, NC 27711
Tel: (919) 541-2296
FTS: 629-2296

Tom Curran
EPA/MDAD
MD-14
RTP, NC 27711
Tel: (919) 541-5351
FTS: 629-5351

Bob Currin
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Harold Crutcher
35 Westall
Asheville, NC 28801
Tel: (919) 253-2539

Robin Davis
EPA/CH/HERL
MD-73
RTP, NC 27711
Tel: (919) 541-2872
FTS: 629-2872

Davis Davis
P.O. Box 12313
RTP, NC 27711
Tel: (919) 549-2333

Robert Denny
EPA/QAB
MD-77
RTP, NC 27711
Tel: (919) 541-2785
FTS: 629-2785

O. L. Dowler
EPA/HERL
MD-56
RTP, NC 27711
Tel: (919) 541-3126
FTS: 629-3126

Ronald Drago
EPA/MDAD
MD-14
RTP, NC 27711
Tel: (919) 541-5486
FTS: 629-5486

Cary Eaton
RTI
P.O. Box 12194
RTP, NC 27709
Tel: (919) 541-6920

Foy W. Edwards
TVA
345 EB
Knoxville, TN 37902
Tel: (615) 632-2071

Susan B. Edwards
NRCD - Air Quality
P.O. Box 27687
Raleigh, NC 27611
Tel: (919) 733-5125

Gardner Evans
EPA/STAB
MD-75
RTP, NC 27711
Tel: (919) 541-2292
FTS: 629-2292

Gary Evans
EPA/STAB
MD-75
RTP, NC 27711
Tel: (919) 541-2294
FTS: 629-2294

B. E. Edmonds
EPA/EMSL
MD-76
RTP, NC 27711

Donald H. Fair
EPA/STAB
MD-75
RTP, NC 27711
Tel: (919) 541-2732
FTS: 629-2732

Bob Faoro
EPA/OAQPS
MD-14
RTP, NC 27711
Tel: (919) 541-5351
FTS: 629-5351

Paul Feder
NIEHS - EBB
P.O. Box 12237
RTP, NC 27709
Tel: (919) 541-5402
FTS: 629-5402

H. L. Fisher
EPA/HERL
MD-74
RTP, NC 27711
Tel: (919) 541-2631
FTS: 629-2631

R. Fisher
EPA/ESRL
MD-80
RTP, NC 27711
Tel: (919) 541-4551
FTS: 629-4551

Nancy Gaskins
RTI
P.O. Box 12194
RTP, NC 27709
Tel: (919) 541-6915

Gerald Gipson
EPA/OAQPS
MD-14
RTP, NC 27711
Tel: (919) 541-5486
FTS: 629-5486

Maurice E. Graves
Northrop Services, Inc.
P.O. Box 12313
RTP, NC 27709
Tel: (919) 549-0411

D. Glover
Rockville International
5529 Chapel Hill Blvd.
Durham, NC 27707
Tel: (919) 942-2407

Bonnee Gryder
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Rel: (919) 541-3080

Ed Hanks
EPA/MDAD
MD-14
RTP, NC 27711
Tel: (919) 541-5474
FTS: 629-5474

F. Hageman
Xonics, Inc
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Martin Hamilton
NIENS - EBB
P.O. Box 12237
RTP, NC 27709
Tel: (919) 541-5402

Tyler Hartwell
RTI
P.O. Box 12194
RTP, NC 27709
Tel: (919) 541-6453

Tom Heiderscheit
EPA/HERL
MD-55
RTP, NC 27711
Tel: (919) 541-2468
FTS: 629-2468

Marvin Hertz
EPA/HERL, MD-56
RTP, NC 27711
Tel: (919) 541-3124
FTS: 629-3124

David O. Hinton
EPA/HERLD
MD-56
RTP, NC 27711
Tel: (919) 541-3126
FTS: 629-3126

Seymour Hochheiser
EPA/EMSL
MD-75
RTP, NC 27711
Tel: (919) 541-2106
FTS: 629-2106

William F. Hunt
EPA/OAQPS
MD-14
RTP, NC 27711
Tel: (919) 541-5351
FTS: 629-5351

R. C. Jordan
Northrop Services, Inc.
P.O. Box 12313
RTP, NC 27709
Tel: (919) 541-2766

Robert B. Jurgens
EPA/ESRL
MD-80
RTP, NC 27711
Tel: (919) 541-4545
FTS: 629-4545

Robert Jungers
EPA/ESRL
MD-78
RTP, NC 27711
Tel: (919) 541-2456
FTS: 629-2456

William E. Klint
NOAA
Fereral Building
Asheville, NC 28801
Tel: (704) 258-2850
FTS: 672-0755

William B. Kuykendal
EPA/IERL
MD-62
RTP, NC 27711
Tel: (919) 541-2557
FTS: 629-2557

Ralph I. Larsen
EPA/ERSL
MD-80
RTP, NC 27711
Tel: (919) 541-4565
FTS: 629-4565

William D. Lee
EPA/QAB
MD-75
RTP, NC 27711
Tel: (919) 541-2293
FTS: 629-2293

Robert E. Lee
EPA/HERL
MD-51
RTP, NC 27711
Tel: (919) 541-2283
FTS: 629-2283

Barry Levene
EPA - Region VIII
1860 Lincoln Street
Denver, CO 80203
Tel: (303) 837-2226
FST: 327-2226

Dan Litton
EPA/HERL
MD-73
RTP, NC 27711
Tel: (919) 541-2873
FTS: 629-2873

Raymond Michie, Sr.
RTI
P.O. Box 12194
RTP, NC 27709
Tel: (919) 541-6492

Randell Morgan
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Gerald K. Moss
EPA/MDAD
MD-14
RTP, NC 27711
Tel: (919) 541-5335
FTS: 629-5335

George C. Murray, Jr.
NCAQ
P.O. Box 27687
Raleigh, NC 27611
Tel: (919) 733-5125

J. E. McCarley, Jr.
EPA/ESED
MD-13
RTP, NC 27711
Tel: (919) 541-5243
FTS: 629-5243

Linda J. McDay
TVA
345-EB
Knoxville, TN 37902
Tel: (615) 632-2071

John S. Nader
EPA/ESRL
MD-46
RTP, NC 27711
Tel: (919) 541-0385

A. Carl Nelson
PEDco
5055 Duke Street
Durham, NC 27701
Tel: (919) 688-6338

William C. Nelson
EPA/HERL
MD-53
RTP, NC 27711
Tel: (919) 541-2330
FTS: 629-2330

W. Norris
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Joan Novak
EPA/ESRL
MD-80
RTP, NC 27711
Tel: (919) 541-4545
FTS: 629-4545

Barbara Nye
EPA/HERL
MD-56
RTP, NC 27711
Tel: (919) 541-3125
FTS: 629-3125

Blaine F. Parr
EPA/HERL
MD-56
RTP, NC 27711
Tel: (919) 541-3123
FTS: 629-3123

C. Don Paulsell
EPA
2565 Plymouth Road
Ann Arbor, MI 48105
Tel: (313) 668-4342
FTS: 374-8342

Debora R. Pizer
EPA/HERL
MD-56
RTP, NC 27711
Tel: (919) 541-3124
FTS: 629-3124

Francis Pooler
EPA/ESRL
MD-59
RTP, NC 27711
Tel: (919) 541-2857
FTS: 629-2857

James Reagan
EPA/ESRL
MD-59
RTP, NC 27711
Tel: (919) 541-4486
FTS: 629-4486

Joan Reece
EPA/HERL
MD-55
RTP, NC 27711
Tel: (919) 541-2466
FTS: 629-2466

Raymond C. Rhodes
EPA/STAB
MD-75
RTP, NC 27711
Tel: (919) 541-2293
FTS: 629-2293

Wilson Riggan
EPA/HERL
MD-54
RTP, NC 27711
Tel: (919) 541-2674
FTS: 629-2674

Charles D. Robson
EPA/HERL
MD-67
RTP, NC 27711
Tel: (919) 541-2625
FTS: 629-2625

Charles E. Rodes
EPA/EMSL
MD-76
RTP, NC 27711
Tel: (919) 541-3076
FTS: 629-3076

Tom Rose
EPA - Region IV
College Station Road
Athens, GA 30601
Tel: (404) 546-3111

Glenn Ross
NCAQ
P.O. Box 27687
Raleigh, NC 27611
Tel: (919) 549-8941

Bill Sensing
EPA/IERL
MD-62
RTP, NC 27711
Tel: (919) 541-2557
FTS: 629-2557

Frank D. Slaveter
EPA
401 M Street, S.W.
EN 340
Washington, DC 20460
Tel: (202) 755-1572

Ben Smith
EPA/IERL
MD-62
RTP, NC 27711
Tel: (919) 541-2557
FTS: 629-2557

Paul E. Smith
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 549-8941

Ralph Sullivan
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 549-8941

Jake Summers
EPA/MDAD
MD-14
RTP, NC 27711
Tel: (919) 541-5395
FTS: 629-5395

José Suñé
EPA/HERL
MD-56
RTP, NC 27711
Tel: (919) 541-3127
FTS: 629-3127

Richard Symonds
Catalytic, Inc.
P.O. Box 240232
Charlotte, NC 28224
Tel: (704) 542-4107

Charles Tate
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

C. E. Tatsch
RTI
P.O. Box 12194
RTP, NC 27709
Tel: (919) 541-5945

Lawrence E. Truppi
EPA/HERL
MD-54
RTP, NC 27711
Tel: (919) 541-2861
FTS: 629-2861

John Van Bruggen
EPA/HERL
MD 55
RTP, NC 27711
Tel: (919) 541-2465
FTS: 629-2465

Darryl von Lehmden
EPA/QAB
MD-77
RTP, NC 27711
Tel: (919) 541-2415
FTS: 629-2415

Betty Wagman
EPA/EMSL
MD-56
RTP, NC 27711
Tel: (919) 541-3125
FTS: 629-3125

Kim Wattenbarger
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

J. E. Whitney
EPA/WA
RD-680
401 M Street, S.W.
Washington, DC 20460
Tel: (202) 426-4477

Cindy Wingarden
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080

Mack Wilkins
EPA/EMSL
MD-45
RTP, NC 27711
Tel-(919) 541-3119
FTS: 629-3119

Marcia Williams
EPA
2565 Plymouth Road
Ann Arbor, MI 48105
Tel: (313) 688-4342
FTS: 374-8323

Max Woodbury
Rockwell International
5529 Chapel Hill Blvd.
Durham, NC 27707
Tel: (919) 493-2471

Chris Woodbury
Xonics, Inc.
P.O. Box 12415
RTP, NC 27709
Tel: (919) 541-3080